



La révolution des données est en marche

Viktor Mayer-Schönberger
Kenneth Cukier

Robert Laffont

111010100010100 101110 00111100010
010101000101011 000101 1000010101010
0111010100010101 010101 0101000101000101
010111 010100 101110 010100 1110100
011000 010101 101010 001011 001010
101010 011100 101010 100010 010101
100010 011011 010100 111000
000101 100010 100110 010101
111010100010101 101011 010100
101111000101010 101110 101000 011101000
110001 010101 010100 010111 100010101
010100 111000 010101 000101 000101010
010101 101010 111011 001011 111101
010100 000101 001110 100010 001010
100110 101110 101010 011101 0101010
0101000101000101 110101 001111000101000111
0001011001100001 011100 01000111010101000
101000101011110 000101 101110101 011

10101110 01011 101000100010 11110
01110101010 010101 011101100010 10010
0111 0001 100010 1010 010111
0101 1000 1010111 0100 1010100
1000 0011 110 010 0101 0001110
1110 0111 1010 0010 1010 1010 010
1010 0010 010 1000 1100 1010 1101
1010 1010 1010101110 0101 101011101
1010 110001010001110 0001 11110001010
001100001010111 1010 1110 0100 1010
0101111010 1000 010 1110 1000 0001

La révolution des données est en marche

Viktor Mayer-Schönberger
Kenneth Cukier

Robert Laffont

VIKTOR MAYER-SCHÖNBERGER
KENNETH CUKIER

BIG DATA

La révolution des données est en marche

*Traduit de l'anglais (États-Unis)
par Hayet Dhifallah*



ROBERT LAFFONT

Ouvrage réalisé sous la direction de Dominique Leglu

Titre original : BIG DATA. A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE,
WORK,
AND THINK

© Viktor Mayer-Schönberger et Kenneth Cukier, 2013

En couverture : © Photo 12 / Alamy

Traduction française : Éditions Robert Laffont, Paris, S.A., 2014

EAN 978-2-221-14451-0

(édition originale : ISBN 978-0-544-00269-2,
published by special arrangement with Houghton Mifflin Harcourt Publishing Company, New York)

Ce document numérique a été réalisé par [Nord Compo](#).

Pour B et v
V. M. S.

Pour mes parents
K. N. C.

1

Maintenant

En 2009, on découvre un nouveau virus de la grippe. Cette souche appelée H1N1, qui associe des éléments issus des virus de la grippe aviaire et de la grippe porcine, se propage rapidement. En l'espace de quelques semaines, les organismes de santé publique du monde entier voient avec inquiétude se profiler la menace d'une terrible pandémie. Certains commentateurs signalent que l'ampleur de l'épidémie risque d'être du même ordre que celle de la grippe espagnole de 1918 qui avait contaminé, elle, cinq cents millions de personnes et avait fait des dizaines de millions de morts. Plus grave encore, aucun vaccin contre le nouveau virus n'est disponible dans l'immédiat. Le seul espoir des autorités sanitaires consiste à ralentir sa propagation. Mais pour cela, il leur faut connaître l'étendue de l'épidémie.

Aux États-Unis, les Centers for Disease Control and Prevention (CDC) ont imposé aux médecins de les tenir informés des nouveaux cas de grippe, pour tenir à jour le tableau de la pandémie. Mais celui-ci s'est révélé toujours en retard d'une ou deux semaines sur la progression réelle de la maladie. À cela, diverses raisons : certains malades ont attendu plusieurs jours après l'apparition des symptômes pour consulter un médecin ; il a fallu compter avec le temps pris pour la transmission des informations aux organisations centrales ; enfin les CDC ne recensaient le nombre de cas qu'une fois par semaine. Pour une maladie à progression rapide, un battement de deux semaines est une éternité. Ce décalage peut complètement bloquer les organismes de santé publique dans les moments les plus cruciaux.

Quelques semaines avant que le virus H1N1 ne fasse ainsi la une de l'actualité, des ingénieurs informaticiens du géant d'Internet, Google,

avaient publié dans la revue scientifique *Nature* un article remarquable sur les responsables de la santé et les informaticiens ; article qui fit sensation mais uniquement dans un cercle restreint de spécialistes. Les auteurs y expliquaient comment Google pouvait « prévoir » la propagation de la grippe hivernale aux États-Unis, non seulement à l'échelle nationale, mais aussi par région et même par État – et cela, en analysant les requêtes des internautes. Sachant que Google fait l'objet de plus de trois milliards de requêtes par jour et les sauvegarde toutes, on comprend que l'entreprise dispose d'une masse considérable de données utilisables.

Google a commencé par établir une liste des 50 millions de mots-clés les plus souvent saisis par les Américains et l'a comparée avec les données des CDC sur la propagation de la grippe saisonnière entre 2003 et 2008. L'idée était de repérer les zones contaminées par le virus de la grippe à partir des requêtes sur Internet. Des tentatives du même ordre et dans le même but avaient déjà eu lieu, qui se fondaient sur les mots-clés saisis sur Internet, mais personne n'avait eu à sa disposition autant de données que Google, ni une telle puissance de traitement ni un si grand savoir-faire statistique.

Pour l'équipe de Google, il était clair que les requêtes visaient à obtenir des informations sur la grippe – au vu du choix de mots-clés tels « médicaments contre la toux et la fièvre ». Cependant, selon elle, là n'était pas la question. Ce qu'il fallait, c'était concevoir un système qui ne prenait pas en compte ce paramètre, mais qui se concentrait uniquement sur la recherche des corrélations entre la fréquence de certaines requêtes et la propagation de la grippe dans le temps et l'espace. Finalement, pour tester les mots-clés, il lui fallut recourir à un nombre colossal de modèles mathématiques différents – 450 millions ! – et comparer leurs prévisions avec les cas de grippe effectifs que les CDC avaient recensés en 2007 et en 2008. Et là, bingo, en plein dans le mille ! Le logiciel trouva 45 mots-clés qui, combinés tous ensemble selon un modèle mathématique, présentaient une forte corrélation entre leur prévision et les chiffres officiels à l'échelle nationale. Tout comme les CDC, ils étaient capables de dire où la grippe s'était propagée mais en comparaison des CDC, ils pouvaient également le faire en temps réel, et non pas une ou deux semaines plus tard.

Ainsi, quand la crise de la grippe H1N1 a éclaté en 2009, le système mis au point par Google s'est révélé un indicateur plus utile et plus précis que les statistiques gouvernementales et leurs inévitables décalages. Les

responsables de la santé publique ont ainsi disposé de renseignements précieux.

Curieusement, ce n'est pas sur des frottis buccaux ou des contacts avec les cabinets de médecins que se base la méthode de Google, mais sur les « big data ». Autrement dit, sur la capacité qu'a cette entreprise à exploiter des informations selon des approches nouvelles pour produire des indications utiles, des biens ou des services extrêmement précieux. Grâce à elle, si une autre pandémie se déclare, le monde disposera d'un meilleur outil pour prédire et donc prévenir sa propagation.

La santé publique n'est que l'un des domaines où les gigantesques masses de données changent énormément la donne. Les big data transforment des pans entiers d'activité. L'achat d'un billet d'avion, par exemple.

En 2003, Oren Etzioni devait aller au mariage de son jeune frère à Los Angeles en partant de Seattle. Des mois avant le jour J, il achète un billet en ligne, convaincu que plus tôt on réserve une place, meilleur marché elle est. Durant le vol, poussé par la curiosité, il demande à son voisin de cabine combien lui a coûté son billet et à quelle date il l'a acheté. Il s'avère que ce dernier l'a payé nettement moins cher qu'Etzioni, alors qu'il l'a acheté bien plus récemment. Furieux, Etzioni pose la même question à plusieurs autres passagers : la plupart ont payé moins cher que lui.

Chez nombre d'entre nous, le sentiment d'avoir été « berné » sur le plan financier se serait dissipé au moment de relever la tablette, de redresser le siège et d'attacher sa ceinture. Mais Etzioni est l'un des informaticiens américains les plus reconnus et pour lui, le monde est une série de problèmes de big data ! Problèmes qu'il sait résoudre depuis l'obtention de son diplôme en informatique à Harvard en 1986.

Professeur à l'université de Washington, avant même que le terme « big data » n'ait été forgé, il a lancé une série de sociétés dont l'activité repose sur l'exploitation d'énormes masses de données. Il a contribué à la création de MetaCrawler, un des premiers moteurs de recherche sur Internet, lancé en 1994 et acquis par InfoSpace, alors importante entreprise en ligne. Il a été cofondateur de Netbot, le premier grand site Web de comparaison de prix, vendu par la suite à Excite. Sa start-up ClearForest, créée pour fournir des solutions d'analyse sémantique de textes, a ultérieurement été acquise par Reuters.

De retour sur la terre ferme, Etzioni est déterminé à trouver un moyen de faire savoir au public si le prix d'un billet d'avion affiché en ligne est une bonne affaire ou non. Une place d'avion est un produit : en principe, sur un même vol, il n'y a aucune différence entre toutes les places et pourtant, leur prix fluctue énormément, en fonction d'une myriade de facteurs pour la plupart connus des seules compagnies d'aviation.

Conclusion d'Etzioni : point n'est besoin de décrypter la raison de ces différences de prix. Il suffit en revanche de prédire si le prix indiqué est susceptible d'augmenter ou de diminuer ultérieurement. La tâche n'est pas facile mais elle est possible : il suffit de soumettre à l'analyse toutes les ventes de billets pour un trajet donné et d'examiner le prix en fonction du nombre de jours avant le départ.

Le prix moyen a-t-il tendance à diminuer ? Il serait judicieux de différer l'achat du billet. Est-ce l'inverse ? Le système devrait recommander d'acheter le billet sur-le-champ au prix indiqué. Autrement dit, Etzioni s'est rendu compte qu'il allait devoir mener à plus grande échelle l'enquête informelle lancée à plus de 9 000 mètres d'altitude. Un problème informatique certes de taille mais soluble. Il s'est donc mis à la tâche.

À partir des informations d'un site Web de voyages sur une période de 41 jours, Etzioni a extrait un échantillon de 12 000 prix, et créé un modèle prédictif permettant aux passagers de réaliser des économies conséquentes. Lors de cette simulation, le modèle n'a pas cherché à comprendre le *pourquoi*, mais seulement le *fait* : ont été ignorées toutes les variables qui ressortissent des décisions des compagnies d'aviation en matière de tarification – nombre de places non vendues, caractère saisonnier ou tarif réduit grâce à la formule magique d'un séjour sur place la nuit du samedi au dimanche. En revanche, la prédiction s'est fondée sur ce qui pouvait être connu : les probabilités recueillies à partir des données relatives à d'autres vols. « Vendre ou ne pas vendre, telle est la question », songea Etzioni. Et du coup, il nomma son projet de recherche Hamlet.

Il évolua vers la création d'une start-up financée par du capital-risque baptisée Farecast. En prédisant que le billet d'avion était susceptible de voir son prix augmenter ou baisser, et de combien, Farecast a offert aux consommateurs la possibilité de choisir à quel moment cliquer sur le bouton « acheter ». Il leur a fourni les informations auxquelles ils n'avaient jamais eu accès auparavant. Dans un souci de transparence, Farecast n'a pas hésité

à évaluer le degré de confiance de ses propres prévisions et à en informer les utilisateurs.

Pour fonctionner, il a fallu fournir au système énormément de données. Afin d'améliorer sa performance, Etzioni s'est procuré l'une des bases de données de réservations de vols provenant du secteur. Muni de ces informations, le système a fait des prévisions fondées sur tous les sièges de tous les vols sur la plupart des itinéraires de l'aviation commerciale américaine en un an. Farecast s'est retrouvé à analyser près de 200 milliards de données concernant le prix des vols, et a fait faire de substantielles économies aux consommateurs.

Avec ses cheveux châtain clair, son large sourire et son visage de chérubin, Etzioni ne semble guère être le genre à priver le secteur des compagnies aériennes de millions de dollars de recettes potentielles. Et pourtant, il a visé encore plus haut. En 2008, il imagine d'appliquer sa méthode à d'autres produits comme chambres d'hôtel, billets de concerts ou voitures d'occasion, qui présentent tous les mêmes caractéristiques : faible différenciation, forte variation des prix et flopée de données disponibles. Avant même que ses projets ne prennent forme, Microsoft est venu taper à sa porte et lui acheter Farecast pour près de 110 millions de dollars avant de l'intégrer au moteur de recherche Bing. En 2012, le système avait atteint un taux de 75 % de prévisions correctes et faisait économiser aux voyageurs 50 dollars par billet en moyenne.

Farecast est l'exemple parfait d'une société prestataire de big data ; elle illustre bien dans quel sens va le monde actuel. Selon ses propres dires, Etzioni n'aurait pu fonder la société cinq ou dix ans plus tôt. Puissance de calcul et volume de stockage nécessaires avaient un coût prohibitif. Les évolutions technologiques ont joué un rôle déterminant dans l'émergence de ce type de société, mais un changement encore plus important y a aussi contribué. Cette subtile évolution concerne le mode d'utilisation des données.

On ne les a plus considérées comme figées ou périmées, voire inutiles une fois atteint le but de leur collecte, par exemple après l'atterrissage de l'avion (ou dans le cas de Google, une fois la requête traitée). Elles se sont transformées en une matière première précieuse, utilisable dans toutes sortes de domaines, un apport vital à la création d'une nouvelle forme de valeur économique. En fait, avec l'état d'esprit approprié, la réutilisation intelligente des données devient source d'innovation et de nouveaux

services : à ceux qui les aborderont avec humilité, qui manifesteront la volonté de les interroger et disposeront des outils adéquats, elles révéleront alors leurs secrets.

Laisser parler les données

Il est facile de repérer les fruits de la société de l'information un peu partout : dans chaque poche un téléphone cellulaire, un ordinateur portable dans chaque sac à dos et de gros systèmes de technologie de l'information dans les back offices. Mais ce qui est moins apparent, ce sont les informations elles-mêmes. Un demi-siècle après que l'utilisation des ordinateurs s'est généralisée dans tous les segments de la société, les données ont commencé à s'accumuler à un point tel que l'on assiste actuellement à un phénomène nouveau et particulier. Non seulement le monde regorge d'informations comme jamais auparavant, mais de surcroît, elles se multiplient rapidement. Le changement d'échelle a entraîné un changement d'état. Le changement quantitatif a conduit à un changement qualitatif. Des sciences comme l'astronomie et la génomique qui, les premières, ont connu l'explosion du phénomène dans les années 2000, ont forgé le terme « big data ». Le concept se répand maintenant dans tous les domaines de l'activité humaine.

Il n'existe pas de définition rigoureuse de ce terme de big data. Au départ, le concept a indiqué qu'à cause d'un volume d'informations devenu trop énorme, les capacités-mémoire des ordinateurs pour effectuer leur traitement étaient dépassées, et les ingénieurs devaient modifier les outils utilisés pour les analyser dans leur totalité. C'est là l'origine de nouvelles technologies de traitement de données comme MapReduce de Google et son équivalent en version open source, Hadoop, créé sous l'impulsion de Yahoo. Elles permettent de gérer de bien plus grandes quantités de données qu'auparavant et – point important – il n'est pas besoin de ranger ces dernières de façon bien ordonnée ou dans des tables de base de données classiques. À l'horizon, se profilent de nouvelles technologies d'analyse de volumes considérables de données qui se passent de hiérarchies rigides et de l'homogénéité d'antan. En parallèle, on peut noter que les sociétés Internet, du fait qu'elles ont été en mesure de collecter de vastes gisements de données et qu'elles ont de sérieuses motivations financières pour les analyser, sont devenues les principales utilisatrices des dernières

technologies de traitement de données. Elles ont ainsi supplanté les sociétés hors ligne qui possédaient, dans certains cas, une expérience de plusieurs décennies.

Aujourd'hui, on peut réfléchir à ce sujet de la manière suivante – et c'est notre démarche dans ce livre : les big data se réfèrent à ce qui peut être accompli à grande échelle et ne peut pas l'être à une échelle plus petite, en matière d'extraction de nouvelles connaissances ou de création de nouvelles formes de valeur, avec comme impact la transformation des marchés, des organisations, de la relation entre les citoyens et les gouvernements, et bien plus encore.

Ce n'est là que le début. L'époque des big data constitue un défi pour notre mode de vie et modifie notre interaction avec le monde. Son plus grand impact ? Quand la société va se rendre compte qu'elle doit mettre un bémol à son obsession de la causalité et se fonder sur de simples corrélations : il ne s'agit plus de connaître le *pourquoi*, mais seulement le *quoi*. Des siècles de pratiques établies vont être ainsi bouleversés. Car la nouvelle démarche remet en question la façon dont nous concevons fondamentalement la prise de décision et l'appréhension de la réalité.

Les big data marquent le début d'une transformation majeure. Comme tant d'autres nouvelles technologies, elles seront à coup sûr les victimes du fameux « Hype Cycle » de la Silicon Valley, le cycle de vie d'une technologie : après avoir fait la couverture des magazines et avoir été portée aux nues lors de conférences organisées par le secteur, la technologie vedette ne suscitera plus d'intérêt. Un certain nombre de start-up dont toute l'activité se sera fondée sur des volumes de données gigantesques connaîtront des difficultés. Mais tant l'engouement que la condamnation révèlent une profonde incompréhension de l'importance du phénomène. Tout comme le télescope nous a permis d'appréhender l'univers et le microscope de comprendre les microbes, les nouvelles techniques de collecte et d'analyse de vastes ensembles de données nous permettront de comprendre notre monde au moyen de méthodes dont nous commençons à peine à mesurer l'importance. Dans ce livre, nous ne voulons pas jouer aux évangélistes mais aux simples messagers du royaume des big data. Sachant que la véritable révolution, on ne le répétera jamais assez, ne réside pas dans les calculs effectués par les machines mais dans les données elles-mêmes et notre façon de nous en servir.

Pour donner la mesure de la révolution en cours dans l'information, examinons les tendances observables dans l'ensemble de la société. Notre univers numérique est en expansion constante. Prenons le cas de l'astronomie. Lorsque le programme Sloan Digital Sky Survey a commencé en l'an 2000, une fois son télescope du Nouveau-Mexique mis en fonctionnement, il a recueilli dans les toutes premières semaines plus de données qu'il n'en avait été amassé durant toute l'histoire de l'astronomie. En 2010, les archives du programme regorgeaient d'un nombre astronomique d'informations – 140 téraoctets ! Mais un successeur, le Large Synoptic Survey Telescope, qui doit être mis en service au Chili en 2016, recueillera cette même quantité de données tous les cinq jours !

On retrouve aussi de telles quantités dans des domaines qui nous touchent de plus près. Lorsque les scientifiques ont décodé le génome humain en 2003, il leur a fallu une décennie de travail intensif pour séquencer les trois milliards de paires de base. Dix ans après, en une seule journée, un seul centre peut séquencer cette même quantité d'ADN. Dans le domaine de la finance, ce sont près de sept milliards d'actions qui changent de mains tous les jours sur les marchés boursiers américains. Pour près des deux tiers, les transactions se font au moyen d'algorithmes informatiques fondés sur des modèles mathématiques qui analysent des montagnes de données afin de prédire les gains tout en essayant de limiter les risques.

Les sociétés Internet ont été particulièrement submergées. Google traite quotidiennement plus de 24 pétaoctets de données^a, ce qui correspond en volume à des milliers de fois la quantité de tous les documents imprimés de la Bibliothèque du Congrès américain. Sur le site Web de Facebook, société qui n'existait pas il y a dix ans, plus de 10 millions de nouvelles photos sont téléchargées toutes les heures. Les membres de Facebook cliquent sur un bouton « J'aime » ou écrivent un commentaire près de trois milliards de fois par jour, créant ainsi une piste numérique que la société peut exploiter pour connaître les préférences de ses utilisateurs. Dans le même temps, les 800 millions d'utilisateurs par mois du service YouTube de Google téléchargent plus d'une heure de vidéo par seconde. Le nombre de messages sur Twitter augmente de près de 200 % par an et, en 2012, il a dépassé les 400 millions de tweets par jour.

Des sciences au secteur de la santé, de la banque à l'Internet, les secteurs varient mais tous racontent la même histoire : la quantité de données à travers le monde augmente à grande vitesse, ce qui dépasse non

seulement la capacité de nos machines, mais aussi celle de notre imagination.

Beaucoup ont tenté d'évaluer la quantité d'informations environnantes et leur taux de croissance, mais avec un succès variable dû au fait que leurs mesures recouvraient des choses différentes. L'une des études les plus approfondies a été réalisée par Martin Hilbert de l'Annenberg School for Communication and Journalism à l'université de Californie du Sud. Il s'est employé à mettre un chiffre sur tout ce qui relève de la production, du stockage, de la communication – livres, peintures, courriels, photographies, musique et vidéo (analogique et numérique) mais aussi jeux vidéo, communications téléphoniques et même systèmes de navigation routière et lettres envoyées par courriel. Il y a inclus des médias comme la télévision et la radio, en se fondant sur le taux d'audience.

À la connaissance de Hilbert, il existait en 2007 plus de 300 exaoctets^b de données stockées. Pour mieux se représenter un tel volume, voici un exemple parlant : un film numérique long métrage peut être compressé en un fichier d'un gigaoctet et un exaoctet équivaut à un milliard de gigaoctets. En bref, c'est énorme. Fait intéressant, en 2007 environ 7 % seulement des données demeuraient analogiques (papier, livres, tirages photographiques, etc.). Le reste était numérique. Mais il n'y pas si longtemps, le paysage était très différent. Même si les idées de « révolution de l'information » et d'« ère numérique » étaient dans l'air depuis les années 1960, ce n'est que plus récemment qu'elles sont devenues réalité. En 2000, le stockage sous forme numérique concernait un quart seulement des informations au niveau mondial. Le reste demeurait stocké soit sur papier, ou sous forme de film, de disques vinyles 33 tours, de cassettes à bande magnétique et autres.

La masse d'informations numériques n'était pas énorme à cette époque – contrairement à ce que pouvaient imaginer ceux qui naviguent sur la Toile et achètent des livres en ligne depuis longtemps. (En fait, en 1986 près de 40 % de la puissance mondiale de calcul permettant d'effectuer des tâches générales étaient assurées par des calculatrices de poche, plus que tous les ordinateurs personnels d'alors.) Mais du fait de l'expansion aussi rapide des données numériques – elles font plus que doubler tous les trois ans, selon Hilbert – la situation s'est rapidement inversée. Les informations analogiques n'ont en revanche que très peu augmenté. Résultat : en 2013, il y aurait eu environ 1 200 exaoctets d'informations stockées dans le monde, dont moins de 2 % sous forme non numérique.

Il est extrêmement difficile de se faire une représentation claire d'un tel volume de données. Seraient-elles sous forme de livres imprimés qu'elles couvriraient la superficie totale des États-Unis sur 52 strates d'épaisseur. Sous forme de CD-ROM empilés, elles s'étireraient jusqu'à la lune en cinq piles séparées. Au III^e siècle av. J.-C., quand Ptolémée II d'Égypte a entrepris de faire stocker un exemplaire de chaque écrit, la Grande Bibliothèque d'Alexandrie représentait la somme de tout le savoir du monde. Avec le déluge numérique qui s'abat désormais sur la planète, à chaque habitant sur Terre correspond aujourd'hui un volume équivalent à 320 fois celui des informations que l'on estime avoir été stockées dans la Bibliothèque d'Alexandrie.

Le rythme s'accélère réellement. La croissance de la quantité d'informations stockées est quatre fois plus rapide que celle de l'économie mondiale, et la puissance de calcul des ordinateurs va neuf fois plus vite. Il n'est pas étonnant que les gens se plaignent de la surabondance d'informations. On est tous abasourdis par ces changements.

Dans une perspective historique à long terme, nous aimerions faire la comparaison entre le déluge actuel de données et une étape antérieure de la révolution de l'information, celle de la presse à imprimer inventée par Gutenberg vers 1439. D'après l'historienne Elizabeth Eisenstein, en cinquante ans, entre 1453 et 1503, près de huit millions de livres ont été imprimés. Plus que tout ce qui avait été produit par les scribes d'Europe, depuis la fondation de Constantinople quelque 1 200 ans auparavant. En d'autres termes, le stock d'informations en Europe a alors presque doublé en cinquante ans, contre trois ans environ aujourd'hui.

Que signifie cette accélération ? Peter Norvig, expert en intelligence artificielle chez Google, s'efforce de nous l'expliquer en faisant une analogie avec les images. Il nous demande de penser dans un premier temps au cheval mythique des peintures rupestres de la grotte de Lascaux, en France, qui datent de l'ère paléolithique il y a environ 17 000 ans. Puis, il s'intéresse à la photographie d'un cheval – ou mieux, aux touches de pinceau d'un tableau de Pablo Picasso, pas bien différentes des peintures rupestres. Quand Picasso vit les images de Lascaux, il eut d'ailleurs cette boutade : « Depuis, nous n'avons rien inventé. »

Picasso n'avait pas tort, mais en un certain sens seulement. Rappelons-nous cette photo du cheval. Alors qu'il fallait un certain temps pour en

dessiner un, il est bien plus rapide aujourd'hui d'obtenir sa représentation avec la photographie. C'est un changement, mais ce n'est sans doute pas le plus essentiel, puisque cela reste fondamentalement l'image d'un cheval. Norvig poursuit : imaginons maintenant que nous fassions la capture de l'image d'un cheval, et que nous la projetions au rythme de 24 images par seconde après l'avoir dupliquée. À ce moment-là, le changement quantitatif a créé un changement qualitatif. Un film est tout à fait différent d'une photographie fixe. Il en est de même pour les big data : en modifiant la quantité, nous modifions l'essence.

Une autre analogie s'inspire des nanotechnologies – domaine où les objets diminuent, au lieu de grossir. Le principe sous-jacent est qu'une fois arrivé au niveau moléculaire, les propriétés physiques changent. Connaître ces nouvelles caractéristiques, c'est devenir capable de concevoir des matériaux permettant d'obtenir des choses impossibles à réaliser jusqu'alors. À l'échelle nanométrique, par exemple, se créent des métaux plus flexibles et des matériaux céramiques étirables. À l'inverse, plus il y a de données, et plus on accède à des résultats nouveaux, ce qui était impossible avec des quantités moindres.

Il arrive que les contraintes auxquelles nous sommes confrontés, et que nous pensons être tout le temps les mêmes, ne dépendent en fait que de l'échelle à laquelle nous opérons. Prenons une troisième analogie, tirée une fois de plus du domaine des sciences. Pour les humains, la loi numéro un de la physique est celle de la gravité : elle régit toutes nos actions. En revanche, au niveau d'un insecte minuscule, la gravité devient quasi immatérielle. C'est une autre loi de l'univers physique qui s'applique, notamment à certains d'entre eux, telles les araignées d'eau : celle de la tension superficielle, qui leur permet de marcher sur l'eau de la mare sans couler.

Avec l'information, comme avec la physique, la taille a son importance. C'est pourquoi Google peut identifier la prévalence de la grippe presque aussi bien que le permettent les consultations effectives de patients, à la base des données officielles. Et ce, en passant au peigne fin des centaines de milliards de mots-clés – une analyse qui livre une réponse quasiment en temps réel, bien plus vite que les sources officielles. De la même manière, la société Farecast fondée par Etzioni peut prédire la volatilité des prix d'un billet d'avion et ainsi avantager le consommateur de façon substantielle. Mais si ces deux entités y parviennent si bien, c'est uniquement parce

qu'elles savent faire l'analyse de centaines de milliards de points de données.

Avec ces deux exemples, on saisit l'importance des big data sur le terrain scientifique et social, et le rôle qu'elles peuvent jouer comme source de valeur économique. On voit comment le monde des big data est appelé à tout bouleverser, depuis les entreprises jusqu'aux sciences en passant par le secteur de la santé, du gouvernement, de l'économie, des sciences humaines et de tous les autres aspects de la société.

Si nous ne sommes qu'à l'aube de l'ère des big data, nous nous en servons pourtant déjà tous les jours. Ainsi, la conception des filtres anti-spams est telle qu'ils soient capables de s'adapter automatiquement avec l'évolution des types de pourriels : le logiciel ne peut pas être programmé pour savoir bloquer le terme « *viagra* » ou ses infinies variantes. Pour mettre en contact des candidats, les sites de rencontres se fondent sur la corrélation qui existe entre leur profil et ses nombreuses caractéristiques et celles des couples dont la rencontre a été précédemment couronnée de succès. Le correcteur orthographique des smartphones suit nos actions et complète son dictionnaire avec des mots nouveaux en fonction de ceux que nous saisissons. Et ces utilisations ne sont qu'un début. Depuis les voitures capables de détecter à quel moment prendre un virage ou freiner jusqu'à l'ordinateur Watson d'IBM capable de battre des êtres humains au jeu télévisé *Jeopardy* !, de nombreux aspects de notre monde vont se métamorphoser.

À la base, les big data sont une affaire de prédictions. Attention à cette description trompeuse qui les fait appartenir à la discipline de l'informatique appelée intelligence artificielle, et plus précisément au champ baptisé apprentissage automatique. Avec les big data, il ne s'agit pas d'essayer d'« apprendre » à un ordinateur à « penser » comme les êtres humains, mais d'appliquer des règles mathématiques à des ensembles de données pour en inférer des probabilités : qu'un message par courriel soit un spam ; que les lettres saisies « *els* » soient en fait « *les* » ; que la trajectoire du piéton qui traverse la chaussée au feu rouge et la vitesse à laquelle il marche indiquent qu'il pourra atteindre le trottoir à temps – il suffit à la voiture automate de ralentir légèrement. Si ces systèmes sont capables de donner de bons résultats, c'est parce qu'ils fondent leurs prédictions sur l'exploitation de grands volumes de données. En outre, ils sont conçus pour s'améliorer, en notant quels sont les signaux et

caractéristiques les plus pertinents au fur et à mesure qu'ils reçoivent de nouvelles données.

À l'avenir – et plus tôt qu'on ne croit – ce qui est aujourd'hui du seul ressort de l'appréciation humaine sera complété ou remplacé par des systèmes informatiques. Cela ne concerne pas seulement la conduite automobile ou la recherche du partenaire idéal, mais aussi des tâches plus complexes. Amazon n'est-il pas capable de recommander le livre idéal, Google de classer les sites Web selon leur pertinence, Facebook de savoir ce qui nous plaît grâce à sa fonction « J'aime » et LinkedIn de deviner ceux que nous connaissons ? Les mêmes technologies s'appliqueront pour diagnostiquer des pathologies, prescrire des traitements et, qui sait, identifier des « criminels » avant même qu'ils n'aient commis un crime. Tout comme Internet a radicalement changé le monde en mettant les ordinateurs en communication les uns avec les autres, les big data elles aussi vont changer la vie dans ses aspects fondamentaux en lui conférant une dimension quantitative qu'elle n'avait jamais eue auparavant.

Plus, désordonnées, bien assez

Les ensembles de données constitueront une source nouvelle de valeur économique et d'innovation. Mais il n'y a pas que ces enjeux. Avec l'analyse des big data et de toutes leurs informations, ce sont la compréhension et l'organisation de notre société qui vont évoluer en trois étapes.

Le premier changement est décrit dans le chapitre 2. Dans ce monde nouveau, nous sommes capables d'analyser des quantités bien plus grandes de données. Dans certains cas, nous pouvons même traiter *toutes* celles relatives à un phénomène particulier. Depuis le XIX^e siècle, la société a été obligée d'avoir recours à des échantillons lorsqu'elle avait affaire à de grands nombres. Procéder par échantillonnage était un artefact nécessaire à une période où les informations étaient rares, et coïncidaient avec la contrainte inhérente à l'analogique. Avant la généralisation de technologies numériques performantes, l'échantillonnage ne nous apparaissait pas comme une forme d'entrave artificielle – en règle générale, il allait de soi. Or, utiliser toutes les données nous donne accès à des détails que nous n'aurions jamais pu voir lorsque nous ne disposions que de quantités plus limitées. Nous pouvons avoir une vision particulièrement claire de sous-

catégories et de sous-marchés impossibles à évaluer du temps des échantillons.

Nous pouvons également être moins exigeants sur l'exactitude des données : c'est le second changement induit par l'examen d'ensembles de données, décrit dans le chapitre 3. Il s'agit d'un compromis : si moins d'erreurs proviennent de l'échantillonnage, alors nous pouvons accepter plus d'erreurs de mesure avec les big data. Si nous ne pouvons faire que peu de mesures, nous ne retenons que les plus importantes. Il est tout à fait justifié de chercher à obtenir le nombre exact. Il est parfaitement inutile de vendre du bétail si le vendeur ne sait pas exactement combien de bêtes compte le troupeau : 100 ou seulement 80 bêtes. Jusqu'à il y a peu, tous nos outils numériques étaient fondés sur la notion d'exactitude : nous partions du principe que les bases de données avaient des moteurs de recherche qui permettaient de trouver les documents répondant parfaitement à notre requête, à la manière des tableurs qui calculent les chiffres dans une colonne.

Ce mode de raisonnement était valable dans un contexte de « petites données » : avec peu de choses à mesurer, nous devions traiter avec le maximum de précision possible ce que nous tenions à quantifier. À certains égards, c'est évident : si une petite boutique peut faire le compte du tiroir-caisse à la fin de la journée jusqu'au dernier centime, nous ne procéderions pas de même – et cela serait d'ailleurs impossible – pour le produit intérieur brut d'un pays. Quand la taille augmente, le nombre d'inexactitudes augmente aussi.

L'exactitude exige de conserver soigneusement les données. Cela peut fonctionner pour de petites quantités, et naturellement certaines situations l'exigent encore : soit on dispose soit on ne dispose pas de suffisamment d'argent en banque pour libeller un chèque. Mais l'utilisation d'ensembles de données beaucoup plus volumineux, dans un monde de big data, nous permet de renoncer en partie à une exactitude rigide.

Les big data sont souvent en désordre, de qualité variable et distribuées entre d'innombrables serveurs à travers le monde. Avec elles, nous nous satisferons d'une notion générale plutôt que de connaître un phénomène jusqu'au plus petit détail – centimètre, centime, atome. Sans renoncer totalement à l'exactitude, nous nous en détachons. Ce que nous perdons en termes de précision à un niveau micro, nous le gagnons en compréhension au niveau macro.

Ces deux changements nous conduisent au troisième, expliqué dans le chapitre 4 : une prise de recul vis-à-vis de l'éternelle quête de causalité. Nous, êtres humains, sommes conditionnés à rechercher les causes, bien que cette quête difficile nous entraîne parfois dans la mauvaise direction. Dans un monde de big data, en revanche, fini l'obsession de la causalité ; dans les données, nous pouvons découvrir des corrélations qui révèlent des tendances et nous apportent des informations vraiment précieuses, voire innovantes. Les corrélations ne nous diront pas précisément *pourquoi* quelque chose se produit, mais elles nous alerteront sur le *fait* qui se produit.

Dans beaucoup de cas, c'est déjà bien assez. Si des millions de dossiers médicaux électroniques révèlent que les patients atteints d'un cancer constatent une entrée en rémission après la prise d'aspirine et de jus d'orange, la cause exacte de l'amélioration de leur santé est alors moins importante que le fait qu'ils soient encore en vie. De même, si nous pouvons faire des économies en ayant connaissance du meilleur moment d'achat d'un billet d'avion sans avoir le détail du système à l'origine des fantaisies des tarifs aériens, c'est amplement suffisant. Avec les big data, il s'agit du *quoi*, et non du *pourquoi*. Il n'est pas toujours nécessaire de connaître la cause d'un phénomène ; laissons plutôt les données parler d'elles-mêmes.

Avant les big data, nous définissions bien avant de collecter les données le petit nombre d'hypothèses auquel notre analyse allait généralement se limiter. Maintenant, quand nous laissons parler les données, des connexions s'établissent dont nous n'avions jamais soupçonné l'existence. C'est ainsi que certains fonds spéculatifs prédisent la performance du marché boursier en faisant l'analyse des tweets. Amazon et Netflix basent leurs recommandations de produits sur une myriade d'interactions dont les utilisateurs laissent la trace sur leurs sites respectifs. Twitter, LinkedIn et Facebook, tous tracent le « graphe social » des relations de leurs utilisateurs pour en savoir plus sur les goûts de ces derniers.

Bien sûr, voilà des millénaires que les êtres humains analysent des données. L'écriture est née en Mésopotamie antique parce que les bureaucrates avaient besoin de disposer d'un outil efficace pour noter des informations et en garder la trace. Depuis les temps bibliques, les gouvernements ont effectué des recensements à seule fin de recueillir

d'immenses quantités de données sur leurs populations, et depuis deux cents ans les actuaires ont de la même manière collecté de vastes quantités de données relatives aux risques qu'ils voulaient mieux appréhender – ou du moins éviter.

Cependant, il fallait beaucoup de temps et d'argent pour organiser cette collecte puis l'analyse des données à l'époque de l'analogique. Et à chaque nouvelle question, il fallait entamer une nouvelle collecte de données et une nouvelle analyse.

L'avènement de la numérisation a permis de franchir une étape importante vers une gestion plus efficace des données en permettant aux ordinateurs de lire les informations analogiques, facilitant par la même occasion le stockage et le traitement à moindre coût. Un progrès qui a considérablement optimisé l'efficacité : là où il fallait des années pour collecter et analyser les informations, tout peut être effectué en quelques jours ou parfois moins. Mais le changement n'a pas affecté certains autres aspects. Par exemple le paradigme analogique, auquel beaucoup d'analystes ont continué de s'accrocher et qui leur faisait penser que les ensembles de données servaient à des fins spécifiques dont leur valeur dépendait. Nos procédures elles-mêmes ont perpétué cet a priori. Si la numérisation a joué un rôle important dans l'avènement des big data, ce n'est pas l'existence d'ordinateurs qui l'a provoqué.

Le concept présenté dans le chapitre 5 permettra de cerner les mutations en cours : *mise en données*. Il s'agit de recueillir tout et n'importe quelle information – y compris celles que nous ne prenions jamais en considération, par exemple la localisation d'une personne, les vibrations d'un moteur ou les forces s'exerçant sur un pont – et de les transformer en données pour les quantifier. Elles peuvent alors nous servir selon de nouveaux modes, l'analyse prédictive, par exemple – la détection d'un risque de panne dans un moteur d'après son échauffement ou ses vibrations – et ainsi apparaît leur valeur implicite, latente.

Le passage de la causalité à la corrélation transforme l'extraction des informations et l'émergence de la valeur latente des données en véritable chasse au trésor. Et il ne s'agit pas là que d'un seul trésor ! Chaque ensemble de données est susceptible de receler une valeur intrinsèque, cachée, non encore mise au jour, pour laquelle il faut engager la course à la découverte.

Les big data modifient la nature des affaires, des marchés et de la société : cela fait l'objet des chapitres 6 et 7. Au xx^e siècle, la valeur est passée des biens matériels – infrastructures physiques comme les terres, les usines... – aux biens immatériels – les marques, la propriété intellectuelle... Maintenant, ce sont les données qui vont devenir un atout précieux, un apport économique vital et la base de nouveaux modèles d'entreprise. C'est la ressource énergétique de l'économie de l'information. Pour l'instant, les données ne sont que rarement intégrées dans le bilan des entreprises, mais il ne s'agit sans doute que d'une question de temps.

Cela fait un moment déjà que des techniques d'analyse de données existent, mais seuls en ont disposé les agences d'espionnage, les laboratoires de recherche et les plus grandes entreprises mondiales. Walmart et CapitalOne ont ainsi été les premières à faire appel à d'énormes masses de données dans les secteurs de la grande distribution et de la banque, ce qui les a bouleversés tous les deux. Aujourd'hui, beaucoup de ces outils se démocratisent (mais pas les données).

C'est au niveau des individus que les big data peuvent provoquer le plus grand choc. Dans un monde où la probabilité et la corrélation priment, l'expertise dans des domaines spécifiques perd de son importance. Dans le film *Le Stratège*, on voit comment, dans le secteur du baseball, les découvreurs de nouveaux talents sont écartés au profit d'une méthode statistique et comment le flair cède la place à des analyses sophistiquées. Les spécialistes ne disparaîtront sans doute pas, mais il leur faudra composer avec les résultats de l'analyse de gros volumes de données. Les conceptions classiques de la gestion, de la prise de décision, des ressources humaines et de l'éducation vont devoir être ajustées.

La plupart de nos institutions voient leur création reposer sur l'idée que le fondement des décisions humaines tient à un petit nombre d'informations exactes et de nature causale. Tout change quand les données deviennent massives, peuvent être traitées avec rapidité et ont une certaine tolérance à l'inexactitude. Les décisions peuvent alors être prises par des machines et non plus par des êtres humains. Nous examinons ce côté sombre des big data dans le chapitre 8.

Voilà des millénaires que la société cumule les expériences permettant de mieux comprendre l'humain et de surveiller son comportement. Mais comment réglementer un algorithme ? Dans les débuts de l'informatique,

les dirigeants politiques ont compris que la technologie pouvait porter atteinte à la vie privée. Depuis lors, la société a établi une réglementation de façon à protéger les informations personnelles. Mais à l'ère des big data, ces lois constituent une ligne Maginot totalement inutile. C'est de leur plein gré que les utilisateurs d'Internet communiquent des informations en ligne – élément central qui fait partie intégrante des services et sans que ces utilisateurs puissent se prémunir contre ce qui les rend vulnérables.

En tant qu'individus, un autre danger nous guette : le déplacement des problématiques du domaine de la vie privée à celui de la probabilité. Les algorithmes prédiront pour un individu la probabilité qu'il ait une crise cardiaque (et donc une prime d'assurance maladie plus forte), celle du non-paiement d'un prêt hypothécaire (avec le risque de se voir refuser un prêt), ou celle de commettre un crime (et d'être éventuellement arrêté préventivement). Ces considérations nous amènent à nous interroger sur le plan éthique : quel rôle pour le libre arbitre ? Ou bien va-t-on s'acheminer vers une dictature des données ? La volonté individuelle devrait-elle prendre le pas sur les big data, même si les statistiques disent le contraire ? Tout comme la presse papier avait préparé le terrain pour des lois garantissant la liberté d'expression – inexistantes auparavant parce qu'il y avait très peu d'expression sous forme écrite à protéger –, il va falloir édicter de nouvelles règles pour l'ère des big data, aux fins de préservation du caractère sacré de la personne.

À bien des égards, notre façon de contrôler et de gérer les données devra changer. Nous entrons dans un monde de prédictions, dans lequel nous risquons de devenir incapables d'expliquer les raisons de nos décisions. Que se passerait-il si demain un médecin ne devait décider d'une intervention médicale qu'à la condition de s'appuyer sur un diagnostic issu d'un ensemble de données concernant le patient ? La « cause probable », norme du système judiciaire, devra-t-elle se transformer en « cause probabiliste » – et dans ce cas, qu'est-ce que cela implique pour la liberté et la dignité humaines ?

Avec l'ère des big data, de nouveaux principes sont requis, que nous exposons dans le chapitre 9. Il ne suffira pas de simplement actualiser d'anciennes règles pour les adapter à de nouvelles circonstances – les valeurs ayant été élaborées dans un monde de « petites données » – il y aura nécessité d'élaborer des principes totalement nouveaux.

On voit d'ici la multitude d'avantages qui se profile pour la société. En effet, les quantités massives de données constituent en partie la solution à de graves problèmes mondiaux : défi du changement climatique, éradication des maladies, amélioration de la gouvernance et développement économique. Mais l'ère des big data nous oblige aussi à autre chose : mieux nous préparer aux changements de nos institutions induits par cette nouvelle technologie et aux transformations que nous aurons nous-mêmes à subir.

Dans la quête de l'humanité pour quantifier et comprendre le monde, les big data marquent une étape importante. Ce qu'il était jadis impossible de mesurer, stocker, analyser et partager fait actuellement l'objet, pour une bonne part, d'une mise en données. Avec l'exploitation de vastes quantités de données et non d'une portion congrue, et le fait de privilégier la masse à l'exactitude, la compréhension des phénomènes va changer. La société va être conduite à tirer parti de la corrélation et à renoncer à sa préférence traditionnelle pour la causalité.

Les big data remettent en question cette illusion gratifiante d'un idéal consistant à identifier les mécanismes de causalité. Une fois encore, nous nous trouvons dans cette impasse historique où « Dieu est mort ». Les anciennes certitudes vont basculer. Ironie de l'histoire, elles sont remplacées, cette fois, par des données plus probantes. De multiples questions se posent : quelle part reste-t-il à l'intuition, à la foi, à l'incertitude, à la latitude d'agir en contradiction avec les preuves, et à l'apprentissage par l'expérience ? Au moment où le monde bascule de la causalité à la corrélation, comment pouvons-nous avancer d'une façon pragmatique, sans saper les fondements mêmes de la société, de l'humanité et le progrès fondé sur la raison ? Ce livre se propose d'expliquer où nous en sommes, de découvrir comment nous y sommes arrivés et de fournir un bon guide qui expose les avantages et les dangers à venir.

^a. 24 millions de milliards d'octets ou environ 200 millions de milliards de bits. (Toutes les notes de bas de page sont du traducteur. Les notes de l'auteur se situent en fin d'ouvrage.)

^b. « Exa » est le préfixe signifiant milliard de milliards, « giga » correspondant à milliard.

2

Plus

Avec les big data, toute la démarche consiste à voir et à comprendre les relations à l'intérieur des éléments d'information et entre ces éléments ; ce que, jusqu'à très récemment, nous avons du mal à appréhender complètement. Jeff Jonas, l'expert en big data d'IBM, affirme que l'on doit laisser les données « vous parler ». D'un certain point de vue, cela peut sembler un peu trop simple. Il y a longtemps que les êtres humains font appel aux données pour découvrir le monde, que ce soit, au sens informel, la myriade d'observations que nous faisons chaque jour ou bien, surtout au cours de ces deux derniers siècles, au sens formel d'unités quantifiées manipulables à l'aide de puissants algorithmes.

L'ère numérique a sans doute rendu plus facile et plus rapide le traitement des données, le calcul de millions de nombres en un clin d'œil, mais lorsque nous évoquons l'idée de données qui parlent, nous entendons par là quelque chose de plus – et quelque chose de différent. Comme précisé dans le chapitre 1, les big data concernent trois grands changements d'état d'esprit qui sont liés et, par conséquent, se renforcent mutuellement. Le premier est le fait que nous soyons capables d'analyser de grands ensembles de données se rapportant à tel ou tel sujet au lieu de nous retrouver contraints de nous accommoder de bien plus petits ensembles. Le deuxième concerne la prise en compte du désordre bien réel des données et non de privilégier leur exactitude. Le troisième est un respect croissant pour les corrélations plutôt que pour la quête perpétuelle d'une insaisissable causalité. Ce chapitre examine le premier de ces changements : l'utilisation de toutes les données disponibles au lieu de se cantonner à une petite portion d'entre elles.

Nous sommes confrontés depuis un certain temps au défi de traiter avec précision d'énormes masses de données. Durant la plus grande partie de notre histoire humaine, nous n'avons travaillé qu'avec une petite quantité de données parce que nous ne disposions que d'outils limités pour les collecter, les organiser, les stocker et les analyser. Nous avons fait le tri des informations sur lesquelles s'appuyer pour n'en garder que le strict minimum et en faciliter ainsi l'examen. C'était une forme inconsciente d'autocensure : nous avons traité la difficulté qu'il y avait à interagir avec les données comme un état de fait regrettable au lieu de la considérer pour ce qu'elle était, à savoir une contrainte artificielle imposée par la technologie de l'époque. Aujourd'hui, le contexte technique a opéré un changement à 179 degrés. Une contrainte demeure, et elle existera toujours, celle du nombre de données que nous sommes capables de gérer, mais la limite est bien moindre qu'auparavant et elle diminuera encore avec le temps.

D'une certaine manière, nous n'avons pas encore pleinement pris la mesure de notre nouvelle liberté de collecter et d'utiliser de plus grands ensembles de données. Deux facteurs ont induit cette idée d'une disponibilité limitée des informations : une bonne partie de notre expérience et la façon dont nos institutions sont conçues. Nous avons estimé que nous ne pouvions collecter qu'une petite quantité d'informations, et c'est ce que nous avons pris l'habitude de faire, concrètement. Nous avons même mis au point des techniques pour utiliser le minimum possible de données. Un des buts des statistiques est, après tout, d'obtenir les résultats les plus pertinents en faisant appel à la plus petite quantité de données possible. En effet, nous avons codifié la limite pratique de la quantité d'informations à utiliser dans nos normes, nos procédures et nos structures incitatives. Si l'on veut mieux saisir le sens du passage au big data, il faut remonter loin dans le temps.

Ce n'est que depuis récemment que des sociétés privées, et même à présent des individus, collectent et trient des informations sur une très grande échelle. Auparavant, la tâche incombait aux institutions les plus puissantes comme l'Église et l'État, ce qui, dans de nombreuses sociétés, revenait au même. La trace la plus ancienne de comptage remonte à environ 5000 av. J.-C. : les commerçants sumériens utilisaient de petites billes d'argile pour indiquer le nombre de marchandises à livrer ou à stocker. Mais compter sur une plus grande échelle relevait des compétences de

l'État. Depuis des millénaires, les gouvernements se sont efforcés de garder la trace de leur population grâce à la collecte d'informations.

Prenons le cas du recensement. Les anciens Égyptiens auraient effectué des recensements, tout comme les Chinois. Il en est question dans l'Ancien Testament, et le Nouveau Testament nous dit qu'un recensement imposé par César Auguste – « En ce temps-là parut un édit de César Auguste, ordonnant un recensement de toute la terre » (Luc 2:1) – fut à l'origine de la fuite de Joseph et Marie vers Bethléem, où naquit Jésus. Un des trésors britanniques les plus vénérés, le *Domesday Book* de 1086, fut à l'époque un inventaire exhaustif et sans précédent de la population anglaise, de ses terres et de ses biens. Des officiers royaux parcoururent l'ensemble du pays en compilant des informations destinées à être consignées dans ce livre – qui prit plus tard le nom de *Domesday*, ou *Doomsday*, parce que le processus ressemblait au Jugement dernier dans la Bible, où la vie de chacun est passée au crible.

Effectuer des recensements coûte cher à la fois en temps et en argent ; le roi William I^{er}, qui commanda l'inventaire, ne vécut pas assez longtemps pour voir le *Domesday Book* finalisé. Le seul moyen de s'éviter pareil fardeau aurait été de renoncer à l'entreprise. De surcroît, en dépit de tout le temps passé et de tout l'argent dépensé, les informations n'étaient qu'approximatives, car il était impossible au recenseur de faire un comptage vraiment parfait. Le terme lui-même de « recensement » vient d'ailleurs du latin *censere*, qui signifie « estimer ».

Il y a plus de trois cents ans, un mercier londonien du nom de John Graunt eut une idée. Il voulait connaître la population de Londres au moment de la peste. Plutôt que de compter chaque individu, il conçut une méthode – nous l'appellerions aujourd'hui « statistique » – qui lui permit de mesurer par *déduction* la taille de la population. Approximative, elle se fondait sur l'idée qu'à partir d'un petit échantillon, il était possible d'extrapoler pour obtenir des informations utiles sur l'ensemble de la population. Même si le procédé employé a son importance, Graunt a simplement décidé de tirer des conclusions générales à partir de son échantillon.

Son système fut largement salué, bien qu'on ait compris par la suite que ses résultats raisonnables n'étaient que pur hasard. Durant des générations, l'échantillonnage demeura entaché de grossières erreurs. Ainsi, pour les recensements et les entreprises similaires à partir de gros volumes de

données, c'est la méthode dite de « force brute », consistant à essayer de compter absolument tous les nombres, qui était de règle.

Du fait de leur complexité, de leur coût et du temps qu'ils nécessitaient, les recensements n'étaient effectués que très rarement. Chez les Romains, qui longtemps se vantèrent de compter une population de centaines de milliers de personnes, un recensement avait lieu tous les cinq ans. Suivant la Constitution des États-Unis, il y eut ordre d'en effectuer un tous les dix ans, étant donné la forte croissance de la population, qui se montait déjà à plusieurs millions. À la fin du XIX^e siècle, cela se révèle problématique ; la masse de données se met à largement dépasser la capacité de suivi du Bureau américain du recensement.

Il faudra huit longues années avant que le recensement entamé en 1880 ne s'achève. Les informations sont caduques avant même d'être disponibles. Pis encore, les responsables estiment qu'il faudrait 13 bonnes années pour que les résultats du recensement de 1890 soient compilés – situation ridicule, sans compter la violation de la Constitution que cela représenterait. Reste qu'il demeure essentiel d'obtenir un décompte de la population à la fois exact et en temps voulu, de façon à répartir l'impôt de façon juste ainsi que d'organiser la représentation de la population au sein du Congrès.

Le Bureau américain du recensement se retrouve en butte aux graves difficultés auxquelles se heurtent aussi les scientifiques et les hommes d'affaires : leurs outils habituels sont dépassés, il faut de nouvelles techniques. Dans les années 1880, la situation devient si critique que, pour le recensement de 1890, le Bureau du recensement passe contrat avec l'inventeur américain Herman Hollerith, qui a imaginé d'utiliser des cartes perforées et des tabulatrices.

Résultat : il est possible de compiler les données en moins d'un an contre huit auparavant, un exploit étonnant qui marque le début du traitement automatique des données (et fournit la base de ce qui deviendra plus tard IBM). Mais en tant que méthode d'acquisition et d'analyse de big data, cela revient très cher : chaque habitant des États-Unis doit remplir un questionnaire et les informations doivent être transférées sur une carte perforée, utilisée pour la compilation.

Voilà le dilemme : faut-il utiliser toutes les données, ou juste une petite partie ? Réunir toutes les données sur le sujet à mesurer, quel qu'il soit, est assurément la démarche la plus judicieuse. Mais c'est tout sauf pratique à

grande échelle. Alors quel échantillon choisir ? Selon certains, la solution la plus appropriée consiste à construire de façon délibérée un échantillon représentatif. Mais en 1934, Jerzy Neyman, statisticien polonais, apporte la démonstration qu'une telle approche conduit à d'énormes erreurs. Pour les éviter, la solution est de choisir de manière aléatoire les personnes à inclure dans l'échantillon.

Les statisticiens ont montré qu'une sélection aléatoire et non l'augmentation de la taille de l'échantillon améliore la précision de l'échantillonnage de manière spectaculaire. Même si cela peut paraître surprenant, un échantillon aléatoire de 1 100 observations individuelles sur une question binaire (oui ou non, avec des chances à peu près égales) est remarquablement représentatif de la population entière. Dans 19 cas sur 20, la marge d'erreur ne dépasse pas les 3 %, que la population totale soit une centaine de milliers ou une centaine de millions de personnes. Il est un peu compliqué d'expliquer la raison mathématique qui se cache derrière ce phénomène, mais pour faire court, on peut dire qu'assez rapidement, et passé un certain point, le nombre d'observations a beau augmenter, l'obtention de nouvelles informations ne modifie les chiffres qu'à la marge.

Le fait que le caractère aléatoire ait pris le pas sur la taille de l'échantillon a été une révélation surprenante. Cela a ouvert la voie à une nouvelle méthode de collecte d'informations. Il est devenu possible, pour un faible coût, de réunir des données en recourant à des échantillons aléatoires, et néanmoins d'extrapoler sur tout l'ensemble, avec des résultats d'une grande précision. C'est ainsi que les gouvernements se sont mis à effectuer tous les ans une version en réduit du recensement, plutôt qu'un recensement tous les dix ans. Et c'est ce qu'ils continuent à faire. Le Bureau américain du recensement, par exemple, réalise plus de deux cents enquêtes économiques et démographiques par an à partir d'échantillons, en plus du recensement décennal qui essaie de compter toute la population. L'échantillonnage s'est révélé une solution au problème de surcharge informationnelle à une époque où la collecte et l'analyse de données étaient très difficiles à réaliser.

Rapidement, cette nouvelle méthode s'est appliquée au-delà des frontières du secteur public et des recensements. L'échantillonnage aléatoire a l'avantage de ramener les problèmes de big data à un problème classique de traitement de données, de dimension gérable. Dans les entreprises, il a été utilisé pour assurer la qualité de la fabrication – améliorer plus

facilement les produits et à moindre frais. Alors que le processus exhaustif de contrôle qualité consistait à examiner un à un chaque produit en bout de ligne de production, il s'est révélé suffisant de contrôler un produit prélevé au hasard dans un lot. La nouvelle méthode par échantillon a également ouvert la voie aux enquêtes de consommateurs dans le secteur de la distribution et aux mini-sondages dans le domaine politique. Une bonne part de ce que nous appelions sciences humaines s'est alors transformée en *sciences sociales*.

Immense succès, l'échantillonnage aléatoire constitue aujourd'hui la clé de voûte de la mesure à grande échelle. Reste qu'il ne s'agit que d'un raccourci, une solution de second ordre par rapport à la collecte et à l'analyse exhaustives des données. Un certain nombre de points faibles sont inhérents à la méthode. Sa précision dépend de la qualité de la sélection aléatoire des données au moment de la constitution de l'échantillon, ce qui n'est pas évident. De graves erreurs dans l'extrapolation des résultats peuvent naître de biais systématiques dans la collecte des données.

On retrouve ce type de problèmes avec les sondages électoraux qui se font par téléphone fixe. L'échantillon désavantage ceux qui n'utilisent que les portables (population plus jeune et plus libérale), comme l'a fait observer le statisticien Nate Silver. D'où un biais qui a conduit à des prédictions électorales erronées. Lors des élections présidentielles de 2008 qui se sont jouées entre Barack Obama et John McCain, les grands instituts de sondage Gallup, Pew, et ABC/Washington Post ont obtenu jusqu'à un voire trois pour cent de différence selon qu'ils incluaient ou pas dans le sondage les utilisateurs de portables – une marge conséquente compte tenu de la lutte serrée entre les deux candidats.

Problème grave, l'échantillonnage aléatoire se prête mal à l'intégration de sous-catégories, étant donné que la ventilation des résultats en sous-groupes de plus en plus petits accroît le risque de prédictions erronées. Ce n'est pas difficile à comprendre. Supposons que vous sondiez un échantillon aléatoire d'un millier de personnes sur leurs intentions de vote lors de la prochaine élection. Si votre échantillon est suffisamment aléatoire, il y a de fortes chances pour que l'opinion de l'échantillon et celle de l'ensemble de la population ne diffèrent au plus que de 3 %. Mais qu'arrive-t-il si plus ou moins 3 % n'est pas suffisamment précis ? Ou qu'on veuille ventiler le groupe en sous-groupes plus petits, selon le sexe, la localisation géographique ou le revenu ?

Et qu'en est-il si on veut cibler une niche dans la population en combinant ces sous-groupes ? Dans un échantillon d'un millier de personnes au total, un sous-groupe de type « riches électriques du Nord-Est » sera largement inférieur à cent personnes. En n'utilisant que quelques dizaines d'observations, la prédiction des intentions de vote de *toutes* les électriques riches du Nord-Est ne pourra être qu'imprécise, même avec une sélection aléatoire proche de la perfection. Et de légers biais dans l'échantillon total conduiront à des erreurs plus importantes au niveau des sous-groupes.

On voit donc que l'exploration approfondie d'une sous-catégorie intéressante des données ne peut se faire par la méthode d'échantillonnage. Ce qui fonctionne au niveau macro ne vaut pas au niveau micro. On peut faire la comparaison avec une épreuve photographique analogique. Vue d'assez loin, elle est impeccable mais, de plus près, les détails deviennent flous.

L'échantillonnage requiert également beaucoup de soin dans la planification et l'exécution. Normalement, il n'est pas possible d'« interroger » à nouveau des données recueillies lors d'un échantillonnage sans avoir bien pensé les questions dès le départ. Donc, même si l'échantillonnage est bien utile en tant que raccourci, ce n'est qu'un raccourci ! Puisque l'ensemble de données est un échantillon et non le tout, il pêche par manque de souplesse et il n'est pas extensible, qualités qui permettraient de le soumettre à une nouvelle analyse, avec un objectif totalement différent de celui pour lequel les données avaient été collectées au départ.

Prenons le cas de l'analyse de l'ADN. En 2012, le séquençage intégral du génome d'un individu frôlait le millier de dollars, un coût qui le rapprochait d'une technique applicable à grande échelle sur un marché de masse. De ce fait, un nouveau secteur émerge. Depuis 2007, la start-up de la Silicon Valley 23andMe offre aux particuliers d'analyser leur ADN pour quelque deux cents dollars. Sa technique permet de révéler les gènes de susceptibilité à développer certaines pathologies comme le cancer du sein ou des problèmes cardiaques. Et en agrégeant les données sur l'ADN de ses clients avec les informations concernant leur santé, 23andMe espère faire de nouvelles découvertes indécélables autrement.

Mais il y a un hic. La société ne séquence qu'une petite portion du code génétique de chacun : elle repère les marqueurs connus de points faibles sur

le plan génétique. Entre ces emplacements précis du génome, des centaines de millions de paires de base d'ADN ne sont pas séquencés. 23andMe ne peut donc répondre qu'aux questions concernant les marqueurs initialement pris en compte. Chaque fois qu'un nouveau marqueur est découvert, l'ADN (ou plus précisément la partie concernée) doit être séquencé à nouveau. Travailler à partir d'un sous-ensemble, au lieu du tout, ne va pas sans compromis : la société trouve ce qu'elle recherche plus rapidement et à moindre coût, mais elle ne peut répondre aux questions qu'elle n'a pas envisagées au préalable.

Le légendaire patron d'Apple, Steve Jobs, a adopté une démarche totalement différente dans sa lutte contre le cancer. Il est devenu l'une des premières personnes au monde à faire séquencer la totalité de son ADN ainsi que celui de sa tumeur. Pour ce faire, il a payé une somme à six chiffres – plusieurs centaines de fois le prix demandé par 23andMe. En contrepartie, il a reçu non pas un échantillon avec un simple ensemble de marqueurs, mais un fichier de données contenant son code génétique dans sa totalité.

Pour choisir un traitement anti-cancer chez un patient lambda, les médecins ne peuvent qu'espérer que l'ADN du malade est suffisamment proche de celui des patients qui ont participé aux essais cliniques. L'équipe médicale de Steve Jobs a pu, elle, faire le choix de thérapies dont les résultats correspondaient au mieux à sa constitution génétique spécifique. Lorsqu'une mutation du cancer le rendait résistant au traitement qui perdait alors en efficacité, les médecins ont pu changer de médicament – « sauter d'un nénuphar à l'autre », selon la formule de Jobs. « Je serai soit un des premiers à pouvoir vaincre un cancer de cette façon soit un des derniers à en mourir », disait-il non sans humour. C'est malheureusement le versant sombre de sa prédiction qui s'est réalisé, mais cette méthode – disposer de toutes les données, pas seulement d'une petite quantité – lui a permis de rester en vie quelques années de plus.

D'une partie au tout

C'est à cause des contraintes sur le traitement de l'information, à une époque où les outils manquaient pour faire l'analyse des données collectées sur le monde, qu'est né l'échantillonnage. C'est un vestige de cette période. Aujourd'hui, les insuffisances en termes de calcul et de compilation sont

moindres. Capteurs, GPS intégré dans les téléphones cellulaires, clics sur Internet et Twitter collectent des données passivement ; les ordinateurs peuvent effectuer de gigantesque calculs avec une facilité accrue.

Maintenant que nous avons les moyens d'exploiter de vastes quantités de données, le concept d'échantillonnage n'a plus de sens. La technique de gestion des données a déjà énormément changé, mais ce sont nos méthodes et notre tournure d'esprit qui prennent plus longtemps à s'adapter.

On observe cependant un changement dans de nombreux domaines : de la collecte de quelques données, on passe à celle du maximum de données possible et, si cela est réalisable, de toutes les données : $N = \text{tous}$.

La formule $N = \text{tous}$ signifie que nous pouvons examiner les données en détail ; une capacité que les échantillons ne nous donnent pas aussi bien, comme nous l'avons vu. Revenons sur l'exemple d'échantillonnage que nous citons ci-dessus : la marge d'erreur ne dépassait pas les 3 % quand nous extrapolions à l'ensemble de la population, chose acceptable dans certaines situations. Mais on a vu que l'on perd les détails, la granularité, c'est-à-dire la capacité à examiner de plus près certains sous-groupes. Une distribution normale n'est, hélas, que normale. Il arrive souvent dans la vie que les choses vraiment intéressantes se trouvent à des emplacements que les échantillons ne parviennent pas pleinement à repérer.

C'est pourquoi « Google Suivi de la grippe » ne s'appuie pas sur un petit échantillon aléatoire mais sur les milliards de requêtes sur Internet aux États-Unis. L'utilisation de toutes ces données permet d'affiner l'analyse jusqu'à prédire la propagation de la grippe dans une ville et pas seulement dans un État ou dans tout le pays. Oren Etzioni de Farecast a utilisé au départ 12 000 points de données, un échantillon, et cela a bien fonctionné. Mais quand Etzioni a commencé d'ajouter encore plus de données, ses prédictions ont été encore meilleures. Farecast a fini par utiliser de surcroît les données sauvegardées concernant la plupart des vols sur les lignes intérieures durant toute une année. « Ce sont des données temporelles – vous en poursuivez la collecte sur le long cours et ce faisant, vous arrivez progressivement à vous faire une idée de plus en plus précise des tendances », déclare Etzioni.

Voilà pourquoi il est souvent préférable de troquer le raccourci que constitue l'échantillonnage aléatoire contre des données plus exhaustives. Cela exige capacité de stockage et techniques d'analyse de pointe, ainsi que des méthodes de collecte des données faciles et d'un coût abordable. Par le

passé, tout faisait problème. Mais aujourd'hui, le coût a baissé et le puzzle est devenu bien moins complexe. Ce qui jadis n'était à la portée que des plus grandes entreprises l'est maintenant de la plupart des sociétés.

Avec toutes les données, il devient possible de repérer les connexions et les détails qui, autrement, resteraient noyés dans l'immensité des informations. Exemple, la détection de la fraude à la carte de crédit qui s'opère en recherchant des anomalies. L'information intéressante se situe dans des données aberrantes qu'on ne peut identifier qu'en les comparant à la masse des transactions normales. C'est un problème de big data. Et du fait que les transactions par carte de crédit se font de manière instantanée, c'est aussi en temps réel que l'analyse doit normalement être menée.

Xoom est une société spécialisée en transferts d'argent à l'international et elle s'est assurée les services des grands noms du domaine des big data. Elle fait l'analyse de toutes les données associées aux transactions qu'elle se retrouve à gérer. En 2011, le système a sonné l'alarme lorsqu'a été détecté, en provenance du New Jersey, un nombre de transactions effectuées par Discover Card légèrement supérieur à la moyenne. « Il a constaté la récurrence d'un phénomène là où il n'aurait pas dû y en avoir, a expliqué John Kunze, directeur général de Xoom. Prise individuellement, chacune des transactions paraissait légale. Mais il s'est révélé qu'une bande de criminels était derrière. Le seul moyen de repérer l'anomalie passait par l'examen de toutes les données – la méthode de l'échantillonnage n'y serait pas parvenue.

Utiliser toutes les données n'est pas forcément une tâche énorme. Les gros volumes de données ne sont pas nécessairement « gros » dans l'absolu, bien que ce soit souvent le cas. Les prédictions de « Google Suivi de la grippe » se fondent sur des centaines de millions d'exercices de modélisation mathématique utilisant des milliards de points de données. La séquence complète d'un génome humain comporte trois milliards de paires de base. Mais ce n'est pas le nombre absolu de points de données en soi, ni la taille de l'ensemble de données, qui font que l'on classe ces exemples dans la catégorie des big data : cela tient plutôt à l'utilisation tant par « Google Suivi de la grippe » que par les médecins de Steve Jobs de la plus grande quantité possible de données sur la totalité, et non celle du raccourci d'un échantillonnage aléatoire.

La découverte du truage de matchs de sumo, ce sport national japonais, est une bonne illustration du fait qu'utiliser $N = \text{tous}$ ne signifie

pas nécessairement gros volume. Ce sport des empereurs voit régulièrement sa réputation mise à mal par des accusations de combats arrangés, ce qui a toujours été nié avec la plus farouche énergie. Steven Levitt, économiste enseignant à l'université de Chicago, s'est attelé à la recherche de traces de corruption dans les archives de matchs qui se sont déroulés sur plus d'une décennie – il a examiné l'ensemble des matchs. Dans un article scientifique savoureux publié dans l'*American Economic Review* et repris dans son ouvrage *Freakonomics*, Levitt et son coauteur ont expliqué pourquoi il était utile de faire l'examen d'autant de données.

Ils ont traqué les anomalies à travers l'analyse des séances de sumo sur onze ans, soit plus de 64 000 combats. Et ils sont tombés sur une mine ! Il y avait bel et bien eu des matchs truqués, mais pas là où la plupart des gens l'auraient soupçonné. Les matchs joués dans le cadre du championnat étaient plus ou moins susceptibles d'être truqués, mais ce que les données indiquaient, c'est qu'il se passait quelque chose de curieux au cours des matchs de fin de tournoi, ceux qui suscitent le moins d'attention, car les enjeux sont devenus bien moins importants (du fait que les lutteurs n'ont aucune chance de remporter un titre).

L'une des particularités du sumo est que les lutteurs, s'ils veulent conserver leur classement et leur revenu, doivent décrocher une majorité de victoires lors des tournois composés de 15 combats. D'où parfois des intérêts asymétriques, quand un lutteur qui totalise 7 victoires et 7 défaites se retrouve confronté à un adversaire qui en est, lui, à 8 victoires et 6 défaites ou mieux. L'issue du match revêt une grande importance pour le premier et quasiment aucune pour le second. Dans ces conditions, et sur la base d'un simple calcul arithmétique, le lutteur qui a besoin de la victoire a de fortes probabilités de l'obtenir.

Se peut-il que ce dernier se batte avec plus de détermination ? Sans doute. Mais les données ont révélé l'existence d'un autre phénomène : le fait que les lutteurs qui ont le plus intérêt à gagner remportent le match en moyenne 25 % de fois plus souvent. Il est difficile de mettre uniquement sur le compte de l'adrénaline un écart aussi grand. Après analyse fine des données, on a vu que, lors du premier match qui mettait face à face les deux mêmes lutteurs, le perdant du combat précédent avait beaucoup plus de probabilités de l'emporter que lors d'affrontements ultérieurs. Ainsi la première victoire apparaît comme un « cadeau » d'un concurrent à l'autre,

étant donné que, dans le monde très soudé du sumo, un « bienfait n'est jamais perdu » !

De telles informations étaient parfaitement visibles, exposées au regard de tous. Mais n'auraient sans doute pas été révélées au moyen de l'échantillonnage aléatoire des combats et de statistiques de base : sans savoir quoi chercher, il aurait été impossible de savoir quel échantillon utiliser. En revanche, bien plus de données – l'examen de l'ensemble des matchs – ont conduit Levitt et son collègue à leur découverte. Une enquête réalisée à partir de big data fait presque songer à une partie de pêche : rien ne dit au départ qu'il y aura prise ni de *quelle sorte de prise* il s'agira.

On constate au passage que les big data ne se mesurent pas forcément en téraoctets. Dans l'exemple du sumo, l'ensemble de données contenait dans sa totalité moins d'octets qu'une banale photographie numérique. Mais il s'agissait bien d'une analyse de big data, puisqu'elle revenait à examiner bien plus de données qu'il n'y en a dans un échantillon aléatoire classique. Parler de big data, c'est vouloir dire « gros » en termes relatifs plutôt qu'en termes absolus ; autrement dit se rapportant à l'ensemble de toutes les données.

On a compris que l'échantillonnage aléatoire avait été pendant longtemps un bon raccourci. Qu'il avait rendu possible l'analyse de données lorsque leur nombre posait problème, à l'époque pré-numérique. Mais de la même manière qu'on perd une partie des informations lorsqu'on convertit une image ou une chanson numériques en un fichier plus petit, on en perd lorsqu'on travaille sur un échantillon. Disposer de tout (ou presque tout) l'ensemble de données donne bien plus de liberté pour explorer, examiner les données sous différents angles ou approfondir certains de leurs aspects.

Une analogie encore meilleure est celle de l'appareil photographique Lytro, qui n'enregistre pas seulement un plan de lumière, comme le font les appareils photographiques conventionnels, mais les rayons de l'ensemble du champ lumineux, lequel en comporte près de 11 millions. Les photographes peuvent décider plus tard sur quel élément d'une image du fichier numérique ils veulent se concentrer, sans avoir à le faire dès le départ, et cela du fait que toutes les informations ont été collectées (en l'occurrence, tous les rayons du champ lumineux sont retenus). Ainsi, les informations sont « réexploitables » bien mieux que pour une photographie

classique où le photographe doit choisir sa mise au point avant d'appuyer sur l'obturateur.

De la même manière, parce que les big data reposent sur la totalité des informations, ou du moins sur le plus d'informations possible, il est permis de procéder à un examen détaillé ou de tester de nouvelles analyses sans risque de flou. Nous pouvons le faire à des niveaux multiples de granularité. C'est grâce à cela qu'il devient possible de déceler le trucage de matchs dans les combats de sumo, de suivre la propagation du virus de la grippe par région et de lutter contre le cancer en ciblant une portion précise de l'ADN du patient. Ce qui nous permet de travailler avec un degré extraordinaire de clarté.

Assurément, il n'est pas toujours nécessaire d'utiliser toutes les données au lieu d'un échantillon. Nous continuons à vivre dans un contexte de ressources limitées. Mais dans un nombre croissant de cas, l'option se justifie d'utiliser toutes les données disponibles et, contrairement à ce qu'on a connu par le passé, l'opération est désormais réalisable.

L'un des domaines qui connaît un énorme bouleversement avec l'arrivée de $N = \text{tous}$ est celui des sciences sociales. La compréhension des données sociales empiriques n'est plus leur domaine réservé ; l'analyse des big data va maintenant remplacer les spécialistes qui étaient hautement qualifiés en matière d'enquêtes. Les disciplines des sciences sociales reposaient en effet largement sur les études d'échantillonnage et les questionnaires. Mais quand les données peuvent être massivement collectées auprès de gens menant leurs activités habituelles, les anciens biais associés à l'échantillonnage et aux questionnaires particuliers disparaissent. Nous pouvons maintenant réunir toutes sortes d'informations – chose impossible auparavant –, qu'il s'agisse des relations que révèlent les appels sur téléphone portable ou de sentiments qui se dévoilent à travers les tweets. Plus important encore, l'échantillon n'est plus nécessaire.

Le scientifique Albert-László Barabási, une des autorités mondiales en théorie des réseaux, voulait étudier les interactions entre les gens à l'échelle de la population tout entière. Avec ses collègues, il a donc examiné la liste anonymisée des appels sur portable que lui a fournie un opérateur de téléphonie mobile desservant près d'une personne sur cinq d'un pays européen – toute la liste sur une période de quatre mois. C'était la première analyse de réseau effectuée au niveau sociétal, en utilisant un ensemble de données dans le même esprit que $N = \text{tous}$. Travailler à une aussi grande

échelle, examiner tous les appels de millions de personnes pendant une période a révélé toute une série de nouveaux éléments d'information qui ne l'auraient probablement pas été par une autre procédure.

L'équipe a ainsi découvert une chose curieuse : contrairement à ce qui se passe lors d'études de moindre envergure, lorsqu'on supprime du réseau des personnes qui ont de nombreux liens dans leur communauté, le reste du réseau social s'affaiblit mais continue cependant à fonctionner. Par contre, si l'on supprime du réseau ceux qui ont des liens en dehors de leur communauté immédiate, le réseau social se désintègre brutalement, comme si sa structure s'effondrait. Ce résultat de taille était inattendu. Qui aurait pu penser que ceux ayant une flopée d'amis proches avaient un impact bien moins important sur la stabilité de la structure du réseau que ceux ayant des liens avec des personnes plus distantes ? Ce résultat laisse entendre que la diversité au sein d'un groupe et dans la société en général joue un rôle essentiel.

Nous avons tendance à considérer l'échantillonnage statistique comme une sorte de fondement immuable, à l'instar des principes de géométrie ou des lois de la gravité. Mais il s'agit d'un concept vieux d'un siècle, élaboré pour résoudre un problème donné à un moment donné sous des contraintes technologiques données, qui tendent à disparaître. Faire appel à un échantillon aléatoire à l'ère des big data est comme tenir une badine pour conduire une automobile ; nous pouvons continuer à utiliser l'échantillonnage dans certains contextes, mais ce n'est plus – et ce ne sera plus – la méthode prééminente d'analyse de nos grands ensembles de données. De plus en plus, nous allons tendre vers l'utilisation du tout big data.

3

Désordre

Il devient possible d'utiliser toutes les données disponibles dans des contextes de plus en plus divers. Mais cela a un coût. Augmenter le volume des données ouvre la porte à l'inexactitude. Certes, les ensembles de données n'ont jamais été exempts d'erreurs et ont toujours contenu des chiffres approximatifs et des bits endommagés. Mais le principe a toujours été de traiter ces derniers comme des problèmes à éliminer au maximum. Ce que l'on s'est refusé de faire, en revanche, c'est de les considérer comme inévitables et d'accepter de les conserver. C'est en cela que réside l'un des changements les plus fondamentaux qui accompagne aujourd'hui le passage de petits à d'énormes volumes de données.

Dans un contexte où les quantités de données sont faibles, réduire les erreurs et garantir une haute qualité est une obligation essentielle. Du fait que la collecte ne porte que sur une petite quantité d'informations, il faut s'assurer que les chiffres sont enregistrés avec la plus grande exactitude possible. Des générations de scientifiques ont optimisé leurs instruments en ce sens, de façon à obtenir des mesures de plus en plus précises, que ce soit pour déterminer la position des corps célestes ou mesurer la taille d'objets vus au microscope. Dans un monde d'échantillonnage, l'obsession de l'exactitude s'est imposée encore plus fortement. Réaliser la seule analyse d'un nombre limité de points de données a toujours fait courir le risque d'amplifier les erreurs, et d'altérer la précision des résultats globaux.

Au cours de l'histoire, on a entrepris la conquête du monde en en faisant la mesure et c'est de là que sont venues la plupart des plus grandes avancées de l'humanité. La quête de l'exactitude a débuté en Europe au milieu du XIII^e siècle, quand les astronomes et les savants ont décidé de quantifier le temps et l'espace d'une manière toujours plus précise – « la

mesure de la réalité », pour reprendre les termes de l'historien Alfred Crosby.

Que l'on mesure un phénomène et ce dernier allait nous devenir compréhensible, telle était la conviction implicite. Plus tard, on a associé la question de la mesure à la méthode scientifique d'observation et d'explication, à savoir la capacité à quantifier, enregistrer et présenter des résultats reproductibles. « Mesurer les choses, c'est les connaître », énonçait lord Kelvin. Une affirmation qui a fait autorité. « La connaissance, c'est le pouvoir », dit aussi Francis Bacon. En parallèle, les mathématiciens et plus tard actuaires et comptables ont élaboré des méthodes précises de collecte, d'enregistrement et de gestion des données.

Au XIX^e siècle, la France – à l'époque, la nation la plus avancée au plan scientifique – avait mis au point un système d'unités de mesure de l'espace, du temps et de bien d'autres domaines, des mesures définies avec précision ; elle avait commencé à faire adopter par les autres nations les mêmes normes, au point d'établir des prototypes devant servir de base aux mesures, ce que des traités internationaux devaient entériner. Ce fut le point culminant de l'ère de la mesure. À peine un demi-siècle plus tard, dans les années 1920, les découvertes de la mécanique quantique ont définitivement fait voler en éclats le rêve consistant à croire qu'on pouvait tout mesurer, et parfaitement. Mais en dehors d'un cercle relativement restreint de physiciens, cette quête obsessionnelle de la mesure parfaite perdura chez les ingénieurs et les scientifiques. Elle se répandit même dans le monde des affaires et du commerce, à partir du moment où on commença de faire appel aux sciences rationnelles comme les mathématiques et les statistiques.

Or, dans bien des situations nouvelles telles que nous en rencontrons aujourd'hui, il est positif d'ouvrir la voie à l'imprécision – le désordre –, ce n'est plus un défaut. C'est un compromis. S'il y a tolérance à l'erreur et que les normes deviennent plus souples, on peut recueillir des données en bien plus grand nombre. Cela ne signifie pas simplement que « le plus l'emporte sur le peu », mais que, parfois, « le plus l'emporte haut la main ».

On rencontre plusieurs formes de désordre. Le terme peut désigner le simple fait que la probabilité qu'il y ait des erreurs croît au fur et à mesure que l'on ajoute de nouveaux points de données. Multiplier par mille le nombre de relevés des tensions qui s'exercent sur un pont augmente ainsi le risque d'obtenir des relevés erronés. Mais le désordre peut aussi augmenter quand on associe différents types d'informations provenant de diverses

sources, qui ne concordent pas toujours parfaitement. Par exemple, utiliser un logiciel de reconnaissance vocale pour établir une distinction entre les diverses réclamations reçues par un centre d'appel et comparer ces données avec le temps de gestion de l'appel nécessaire à l'opérateur peut donner un aperçu imparfait mais utile de la situation. Le désordre peut également concerner la disparité des formats, qui nécessite un « nettoyage » des données avant leur traitement. Il existe de multiples appellations pour IBM, fait observer par exemple l'expert en big data D. J. Patil : ce peut être IBM ou T. J. Watson Labs ou encore International Business Machines. Et le désordre survient aussi au moment de l'extraction ou du traitement même des données, puisque par cette opération elles subissent une transformation qui les fait devenir quelque chose d'autre. Exemple : quels sont les sentiments qui ressortent d'une analyse de tweets et permettent de prédire les résultats du box-office et les recettes de Hollywood. Le désordre a lui-même son propre désordre.

Supposons que nous ayons à mesurer la température dans un vignoble. Si nous n'avons à notre disposition qu'un seul thermomètre pour l'ensemble du vignoble, qui compte des centaines de pieds de vigne, nous devons nous assurer de sa précision et faire en sorte qu'il fonctionne en permanence : aucun désordre n'est autorisé. En revanche, si nous plaçons un capteur au pied de chaque cep, nous pouvons avoir recours à des appareillages moins chers et moins sophistiqués, tant qu'ils ne provoquent pas un biais systématique. Le risque existe qu'à certains endroits, quelques capteurs affichent des données incorrectes, et fournissent un ensemble de données moins exact, ou plus « déstructuré » que celui obtenu avec un unique capteur précis. Un relevé pourra être inexact, mais les multiples relevés fourniront au total un tableau plus global. Du fait que cet ensemble de données comporte un nombre plus élevé de points de données, son intérêt est bien supérieur et cela compense sans aucun doute son désordre.

Supposons maintenant que nous relevions les résultats des capteurs plus fréquemment. Si nous prenons une mesure par minute, nous pouvons être quasiment certains que les données arriveront dans une séquence parfaitement chronologique. Mais si nous accélérons la fréquence des mesures, avec dix ou cent relevés par seconde, la séquence risque de perdre en exactitude. Parce que les informations circulent en réseau, une valeur peut arriver en retard ou pas dans le bon ordre, voire se perdre dans le flot. Les informations seront légèrement moins précises, mais leur volume leur

confère une telle importance que cela vaut la peine de renoncer à une stricte exactitude.

Dans le premier exemple, nous avons sacrifié à l'abondance des données la précision de chaque point de données et en contrepartie, nous avons pu voir des détails que nous n'aurions pas obtenus autrement. Dans le second, nous avons renoncé à l'exactitude en contrepartie de la fréquence, et en retour nous avons décelé un changement qui nous aurait échappé. Il est possible de dépasser le problème des erreurs, à condition d'y mettre suffisamment de moyens – après tout, la Bourse de New York fonctionne au rythme de 30 000 transactions par seconde, dans un domaine où l'exactitude de chaque séquence est d'extrême importance – mais, dans de nombreux cas, mieux vaut tolérer l'erreur que de s'acharner à la prévenir.

Nous pouvons accepter une certaine dose de désordre en contrepartie du volume. Comme le dit le cabinet conseil en technologies Forrester : « Il arrive parfois que deux plus deux égalent 3,9 et ça peut aller. » Bien sûr, les données ne peuvent pas être totalement incorrectes, mais nous acceptons volontiers de sacrifier un peu de précision pour connaître la tendance générale. Avec les données massives, les chiffres doivent être pris en termes de probabilités plus que de précision. Il nous faudra du temps avant que l'on s'habitue à ce changement qui comporte son lot de problèmes, que nous aborderons plus loin dans le livre. Pour l'instant, il suffit de noter qu'il nous faudra souvent accepter la notion de désordre quand nous changeons d'échelle et passons à un niveau supérieur.

L'évolution est similaire dans d'autres domaines de l'informatique, en plein changement eux aussi. Ainsi, chacun sait que la puissance de traitement n'a cessé d'augmenter au fil des ans en suivant la courbe prédite par la loi de Moore, selon laquelle le nombre de transistors double à peu près tous les deux ans dans une puce électronique. Avec cette amélioration continue, les ordinateurs sont devenus plus rapides et sont dotés d'une capacité-mémoire plus étendue. Ce que nous savons moins, c'est que les algorithmes qui font tourner nos systèmes sont bien plus rapides et performants – et dépassent même la tendance donnée par la loi de Moore. Cependant, gardons en mémoire que c'est le fait de disposer d'un plus grand nombre de données qui engendre les changements majeurs, plus encore que la rapidité accrue des puces ou les meilleures performances des algorithmes.

Ceux utilisés pour le jeu d'échecs, par exemple, ont à peine changé ces dernières décennies, du fait de notre parfaite connaissance des règles du jeu d'échecs et de leur côté très strict. Si les programmes informatiques de jeu d'échecs sont plus performants aujourd'hui, c'est en partie dû au fait qu'ils jouent mieux leur dernière manche. Et ce, parce qu'une plus grande quantité de données a pu alimenter les systèmes. Les dernières manches – lorsqu'il ne reste que six pions ou moins sur l'échiquier – ont en effet été analysées de façon exhaustive et tous les coups possibles ($N = \text{tous}$) représentés sous forme d'un tableau qui, lorsqu'il est décompressé, remplit plus d'un téraoctet (mille milliards d'octets) de données. Le jeu d'échecs électronique peut ainsi jouer la dernière manche sans faute. Aucun être humain ne sera jamais capable de le surpasser.

C'est dans le domaine du traitement du langage naturel qu'il a été magistralement démontré à quel point une plus grande quantité de données l'emportait sur de meilleurs algorithmes : c'est ainsi que les ordinateurs apprennent à analyser les mots tels que nous les utilisons dans le langage courant. Vers l'année 2000, Michele Banko et Eric Brill, chercheurs chez Microsoft, voulaient améliorer le vérificateur de grammaire intégré dans le programme Word de la société. Ils n'étaient pas certains qu'il serait plus utile de consacrer leurs efforts à améliorer les algorithmes existants, ou bien de trouver de nouvelles techniques ou encore d'ajouter des fonctions plus élaborées. Avant de se lancer sur l'une de ces pistes, ils décidèrent de continuer à utiliser les méthodes classiques mais d'y intégrer beaucoup plus de données et de voir ce qui se produirait. La plupart des algorithmes d'apprentissage machine s'appuyaient sur des corpus de textes qui pouvaient totaliser un million de mots. Banko et Brill ont décidé de retenir quatre algorithmes communs et de leur fournir un nombre de données supérieur de trois ordres de grandeur : 10 millions de mots, puis 100 millions et, finalement, un milliard de mots.

Les résultats ont été stupéfiants. L'intégration régulière de quantités supplémentaires de données a remarquablement amélioré la performance de l'ensemble des quatre types d'algorithmes. Encore plus étonnant, l'algorithme, le moins performant quand il se fondait sur un demi-million de mots, obtint les meilleurs résultats quand il en analysa un milliard. Sa précision était passée d'un taux de 75 % à plus de 95 %. À l'inverse, l'algorithme qui fonctionnait le mieux avec une faible quantité de données obtint les plus mauvais résultats avec des quantités plus importantes, même

si, comme les autres, il s'améliorait beaucoup, passant de près de 86 % à près de 94 % en termes de précision. Selon Banko et Brill, « ces résultats indiquent que nous devrions sans doute reconsidérer l'idée d'établir un compromis entre dépenser du temps et de l'argent dans la conception d'algorithmes ou en dépenser dans le développement de corpus de données », comme ils l'ont écrit dans l'un de leurs rapports de recherche.

Ainsi, le plus l'emporte sur le moins. Et parfois, plus l'emporte plus intelligemment. Qu'en est-il alors du caractère déstructuré ? Quelques années après que Banko et Brill eurent intégré toutes ces masses de données, les chercheurs du concurrent Google se sont mis à suivre les mêmes pistes de réflexion – mais à plus grande échelle encore. Au lieu de tester des algorithmes avec un milliard de mots, ils en ont utilisé un billion (mille milliards). Ce n'est pas pour mettre au point un vérificateur de grammaire que Google s'y est attaqué mais pour résoudre un problème encore plus complexe : la traduction dans d'autres langues.

Les pionniers de l'informatique ont caressé le rêve de la traduction dite automatique dès le début dans les années 1940, quand les appareillages étaient constitués de tubes à vide qui occupaient une salle entière. L'idée a pris un caractère d'urgence particulier au cours de la guerre froide, époque où les États-Unis manquaient de ressources humaines pour traduire dans de brefs délais les énormes quantités de documents en russe qu'ils collectaient, écrits et oraux.

Au départ, les informaticiens ont opté pour une combinaison de règles de grammaire et de dictionnaire bilingue. En 1954, un ordinateur IBM pouvait traduire soixante phrases du russe en anglais, en utilisant 250 paires de mots du vocabulaire de l'ordinateur et six règles grammaticales. Les résultats étaient très prometteurs. La phrase en russe « *Mi pyeryedayem mislyi posryedstvom ryechyi* » était entrée dans l'appareil IBM 701 par le biais de cartes perforées, et la phrase suivante en ressortait « Nous transmettons des pensées au moyen du langage ». Les soixante phrases étaient « bien traduites », selon un communiqué de presse d'IBM célébrant l'événement. Le directeur du programme de recherche, Leon Dostert de l'université de Georgetown, prédit que la traduction automatique serait un « fait acquis » dans « cinq, peut-être même trois ans ».

Malheureusement, après ce succès, ce fut l'impasse. En 1966, un comité de pontes de la traduction automatique dut reconnaître que l'entreprise était un échec. Le problème se révélait plus difficile qu'ils ne l'avaient imaginé.

Apprendre aux ordinateurs à traduire ne consiste pas à leur apprendre uniquement les règles, mais aussi les exceptions. La traduction n'est pas une simple affaire de mémorisation et de remémoration ; c'est aussi choisir les mots appropriés parmi plusieurs options. « Bonjour » correspond-il vraiment à « *good morning* » en anglais ? Ou bien à « *good day* », ou « *hello* » ou « *hi* » ? Réponse : cela dépend...

À la fin des années 1980, les chercheurs d'IBM ont eu une nouvelle idée. Au lieu d'essayer d'intégrer dans un ordinateur des règles linguistiques explicites, à la manière d'un dictionnaire, ils ont décidé de laisser l'ordinateur jongler avec les probabilités pour déterminer le mot ou la phrase approprié(e) dans une langue en comparaison avec une autre. Dans les années 1990, le projet Candide d'IBM éplucha les comptes rendus des travaux parlementaires canadiens couvrant une période de dix années publiés à la fois en français et en anglais – soit près de trois millions de paires de phrases. Du fait du caractère officiel des documents, les traductions étaient de très haute tenue. Et selon les normes de l'époque, la quantité de données était énorme. La traduction automatique statistique – nom donné à cette technique – eut son heure de célébrité ; ce qui était un défi en termes de traduction devint un énorme problème mathématique. Et apparemment, cela a fonctionné. Brusquement, la traduction informatique a fait de nets progrès. Toutefois, après le succès de ce bond conceptuel, IBM n'a effectué que de très légères améliorations en dépit de tout l'argent investi. La compagnie a même fini par mettre un terme au projet.

Moins d'une décennie plus tard, en 2006, Google s'est à son tour lancé dans la traduction, dans le cadre de sa mission d'« organiser les informations sur le monde et de les rendre accessibles et utiles au monde entier ». Plutôt que de pages de textes correctement traduites en deux langues, Google a eu recours à un ensemble de données plus large mais également bien plus déstructuré : l'Internet mondial dans son entier et plus encore. Son système a puisé dans les traductions du monde entier, de façon à entraîner l'ordinateur – sites Web d'entreprises dans de multiples langues, traductions de documents officiels et rapports émanant d'organismes intergouvernementaux telles les Nations unies et l'Union européenne... et même traductions de livres issues du projet de numérisation de livres de Google. Là où Candide avait utilisé trois millions de phrases traduites avec soin, le système de Google a exploité des milliards de pages de traductions de qualité très variable, d'après le responsable de Google Traduction, Franz

Josef Och, l'une des autorités les plus éminentes dans le domaine. Dans son corpus de mille milliards de mots, il y avait 95 milliards de phrases en anglais, même si elles étaient de qualité douteuse.

En dépit du caractère déstructuré des données d'entrée, le service de Google est celui qui fonctionne le mieux. Ses traductions sont plus exactes que celles des autres systèmes (même s'il y a toujours des imperfections). Et il est de loin bien plus riche. Au milieu de l'année 2012, son ensemble de données couvrait plus de 60 langues. Il peut même accepter des entrées vocales dans 14 langues et produire des traductions fluides à partir de là. Du fait qu'il traite la langue comme un corpus de simples données déstructurées auxquelles sont affectées des probabilités, il peut même traduire du hindi vers le catalan, pour lesquelles il n'existe que très peu de traductions directes. Dans ces cas-là, il utilise l'anglais comme passerelle. Et sa méthode est bien plus souple que les autres, car il ajoute et retranche des mots au fur et à mesure qu'ils entrent dans l'usage courant ou en disparaissent.

Ce n'est pas parce qu'il utilise un algorithme plus intelligent que le système de traduction de Google fonctionne si bien. Il fonctionne bien parce que ses créateurs, comme Banko et Brill chez Microsoft, lui ont fourni un plus grand volume de données – et pas uniquement des données de haute qualité. Google a pu faire appel à un ensemble de données *des dizaines de milliers* de fois plus grand que celui du projet Candide d'IBM et cela, parce qu'il a décidé de composer avec le désordre. Le corpus de mille milliards de mots que Google a créé en 2006 est le résultat de la compilation du bric-à-brac que contient Internet – des « données à l'état sauvage », en quelque sorte. Il a servi de « trousse de formation » au système pour calculer par exemple la probabilité qu'un mot en anglais en suive un autre. L'ancêtre dans le domaine, le célèbre « Brown Corpus^a » des années 1960, avec son total d'un million de mots anglais, en était loin ! L'utilisation d'un plus grand ensemble de données a permis de réaliser des progrès énormes dans le traitement du langage naturel, sur lequel se fondent les systèmes de reconnaissance vocale et de traduction informatique, par exemple. « Des modèles simples et une grande quantité de données surpassent des modèles plus élaborés fondés sur une quantité moindre de données », ont écrit Peter Norvig, gourou de l'intelligence artificielle chez Google, et ses collègues dans un article intitulé « The Unreasonable Effectiveness of Data^b ».

Comme l'ont expliqué Norvig et ses coauteurs, c'est le désordre qui en est la clé : « À certains égards, ce corpus est un pas en arrière par rapport au Brown Corpus : il provient de pages Web non filtrées et contient donc des phrases incomplètes, des fautes d'orthographe, des erreurs de grammaire et toutes sortes d'autres erreurs. Il n'est pas soigneusement annoté à l'aide d'étiquettes créées manuellement par un étiqueteur. Mais le fait qu'il soit un million de fois plus grand que le Brown Corpus l'emporte sur tous ces inconvénients. »

Le plus l'emporte haut la main

Le désordre est difficile à accepter pour les analystes habitués à un échantillonnage conventionnel et qui, toute leur vie, ont axé leurs efforts sur la prévention et l'éradication du désordre. Ils s'efforcent de réduire les taux d'erreur lors de la collecte des échantillons et ils les testent pour détecter d'éventuels biais avant d'annoncer leurs résultats. Ils font appel à de multiples stratégies pour réduire les erreurs, notamment en veillant au respect d'un protocole exact dans la collecte des échantillons par des experts formés spécialement à cet effet. Cette mise en œuvre est coûteuse même pour un nombre limité de points de données, et difficilement adaptable à des quantités massives de données. Non seulement elles reviennent bien trop cher, mais il est peu probable que la rigueur des normes de collecte soit conservée à une telle échelle. Même sans intervention humaine, le problème ne serait pas résolu.

Passer à un monde de big data nous demandera de changer de point de vue sur les mérites de l'exactitude. Effectuer des mesures dans ce monde numérique et connecté du XXI^e siècle avec un état d'esprit conventionnel serait passer à côté d'un point essentiel. Comme évoqué plus haut, l'obsession de l'exactitude est un artefact de l'époque de l'analogique où les informations étaient peu nombreuses. Quand les données étaient rares, chaque point de données était précieux, et il fallait donc soigneusement éviter d'en laisser un fausser l'analyse.

Aujourd'hui, nous ne vivons pas dans un contexte aussi pauvre en informations. Les ensembles de données sont toujours plus complets, ils contiennent non pas simplement un petit fragment du phénomène disponible, mais parfois la description du phénomène tout entier, et nous n'avons plus besoin de redouter que certaines données viennent fausser

l'analyse générale. Plutôt que de viser à éliminer la moindre trace d'inexactitude à un coût de plus en plus élevé, nous effectuons des calculs qui tiennent compte du facteur désordre.

Voyez comment les capteurs font leur entrée dans les usines. À la raffinerie Cherry Point de BP à Blaine, dans l'État de Washington, des capteurs sans fil sont installés dans toute l'usine, formant un filet invisible qui livre de vastes quantités de données en temps réel. Il fait très chaud et il y a de nombreuses machines électriques, ce qui peut fausser les relevés et faire que le système fournit des données inexactes. Mais ces anomalies sont compensées par l'énorme quantité d'informations enregistrée par les capteurs. La simple augmentation de la fréquence des relevés et de leur nombre peut constituer un bon compromis. En mesurant en permanence la tension exercée sur la tuyauterie plutôt qu'à certains intervalles de temps, la société BP a découvert que certains types de pétrole brut étaient plus corrosifs que d'autres – caractéristique indétectable et donc imparable, quand l'ensemble de données dont elle disposait était moindre.

Avec une quantité de données bien plus considérable et d'un nouveau type, il arrive que l'objectif ne soit plus d'être exact mais de deviner la tendance générale. Le passage à grande échelle modifie non seulement les attentes en termes de précision mais aussi la capacité pratique à obtenir des résultats exacts. Cela peut paraître paradoxal au premier abord, mais le traitement imparfait de données imprécises nous permet de réaliser de bien meilleures prévisions, et donc de mieux comprendre le monde où nous vivons.

Il faut souligner que le désordre n'est pas une caractéristique inhérente aux gros volumes de données. Il est dû à l'imperfection de nos outils de mesure, d'enregistrement et d'analyse des informations. La technologie atteindrait-elle la perfection que le problème de l'inexactitude disparaîtrait. Mais entre-temps, et cela risque de durer, le désordre est une réalité pratique et il faut faire avec. Augmenter l'exactitude ne rime souvent à rien sur le plan économique, étant donné qu'avec des quantités plus grandes de données, la valeur ajoutée s'accroît. À l'instar des statisticiens qui, jadis, se sont désintéressés des échantillons de plus grande taille pour privilégier leur caractère aléatoire, nous pouvons vivre avec une certaine dose d'imprécision en contrepartie d'une quantité plus grande de données.

À cet égard, le Billion Prices Project constitue un exemple intéressant. Tous les mois, l'U.S. Bureau of Labor Statistics^e publie l'indice des prix à la

consommation, ou CPI, à partir duquel on calcule le taux d'inflation. Le chiffre est essentiel pour les investisseurs et les entreprises. La Réserve fédérale, par exemple, le prend en compte quand elle décide d'augmenter ou de diminuer les taux d'intérêt et les entreprises se fondent sur l'inflation pour les augmentations de salaires. Le gouvernement fédéral y fait appel pour indexer les paiements comme les prestations de sécurité sociale et mesurer l'intérêt qu'il paie sur certaines obligations.

Pour obtenir ce chiffre, le Bureau of Labor Statistics emploie des centaines d'employés qui passent des milliers d'appels, envoient des fax et effectuent des visites à des magasins et des bureaux dans 90 villes à travers tout le pays afin de recueillir et communiquer des informations sur près de 80 000 prix de toutes sortes allant de celui des tomates aux tarifs des taxis. La production de cet indice coûte près de 250 millions de dollars par an. Pour une telle somme, les données sont claires et impeccables. Mais au moment où les chiffres sont publiés, ils datent déjà de quelques semaines. Comme l'a montré la crise financière de 2008, quelques semaines peuvent constituer un laps de temps extrêmement long. Il faut aux décideurs un décalage plus court dans l'accès aux informations sur les taux d'inflation afin de mieux y réagir. Chose impossible avec des méthodes classiques axées sur l'échantillonnage et la précision des prix.

Pour pallier ce problème, deux économistes de l'Institut de technologie du Massachusetts (MIT), Alberto Cavallo et Roberto Rigobon, ont mis au point une autre méthode fondée sur les big data, qui emprunte une voie beaucoup plus déstructurée. Avec l'aide d'un logiciel qui parcourt le Web, ils ont collecté le prix de cinq cent mille produits vendus chaque jour aux États-Unis. Les informations sont approximatives, et il n'est pas facile de comparer tous les points de données faisant l'objet de la collecte. Mais en combinant cette réunion massive de données avec une analyse ingénieuse, le projet a pu détecter un virage déflationniste immédiatement après le dépôt de bilan de Lehman Brothers en septembre 2008, alors qu'il a fallu à ceux qui tablaient sur les données officielles du CPI attendre le mois de novembre pour pouvoir le constater.

Du projet du MIT est née une entreprise commerciale, Price-Stats, à laquelle les banques et autres organismes font appel avant de prendre des décisions d'ordre économique. Elle compile les prix de millions de produits vendus quotidiennement par des centaines de détaillants dans plus de 70 pays. Naturellement, la prudence s'impose dans l'interprétation des chiffres,

mais ils indiquent mieux les tendances inflationnistes que les statistiques officielles. Un avantage non négligeable, fondé sur des données en temps réel. (La méthode sert également de mécanisme de contrôle externe crédible des instituts statistiques nationaux. Ainsi, *The Economist*, qui n'a pas confiance dans la méthode de calcul de l'inflation pratiquée en Argentine, préfère s'appuyer sur les chiffres de PriceStats.)

Le désordre en action

Faire le choix d'un plus grand nombre de données, seraient-elles déstructurées, c'est aujourd'hui ce vers quoi on tend, plus que vers un nombre limité de données parfaitement exactes. Et ce, dans de nombreux domaines de la technologie et de la société. Prenons le cas de la mise en catégories de certains contenus. Durant des siècles, les êtres humains ont élaboré des taxonomies et des index pour stocker et retrouver les informations. Ces systèmes hiérarchiques ont de tout temps été imparfaits, comme quiconque en garde le souvenir douloureux, s'il a voulu se familiariser avec le catalogue sur fiches d'une bibliothèque. Mais dans un contexte de petites quantités de données, c'était satisfaisant. Toutefois, si l'échelle change de plusieurs ordres de grandeur, ces systèmes, qui supposent que tous les éléments sont parfaitement rangés, s'effondrent. Par exemple, en 2011, il y avait plus de 6 milliards de photographies provenant de plus de 75 millions d'utilisateurs sur le site de partage de photographies Flickr. Tenter d'étiqueter chaque photo selon des catégories préétablies aurait été peine perdue. Aurait-on créé une catégorie intitulée « Chats ressemblant à Hitler » ?

Au lieu de cela, les taxonomies impeccables sont remplacées par des mécanismes plus flous mais éminemment plus flexibles, qui s'adaptent à un monde changeant voire en mutation. Quand nous téléchargeons des photos sur Flickr, nous leur collons une « étiquette ». Autrement dit, nous leur attribuons un libellé à notre convenance, qui servira à organiser les documents et à faire des recherches. Les étiquettes sont apposées au cas par cas : elles sont créées sans catégories prédéfinies et normalisées, et donc sans taxonomie préexistante à laquelle il faudrait se conformer. Au contraire, n'importe qui peut ajouter de nouvelles étiquettes par une simple saisie du libellé. L'étiquetage, utilisé sur les sites de médias sociaux tels Twitter, les blogs, etc., s'est révélé la norme de facto pour classer du

contenu sur Internet. Il facilite la navigation à travers le vaste contenu de la Toile, en particulier pour des éléments qui ne sont pas fondés sur des textes – images, vidéos et musique – et pour lesquels les requêtes à partir de mots-clés ne fonctionnent pas.

Bien sûr, certaines étiquettes peuvent présenter des fautes d'orthographe, ce qui entraîne une certaine imprécision – non dans les données elles-mêmes mais dans la façon dont elles sont organisées. Et l'esprit traditionnel formé à l'exactitude en souffre. Mais en contrepartie de ce désordre dans l'organisation des collections de photos, les étiquettes gagnent en richesse et, par extension, nous obtenons un accès plus poussé et plus large à nos images. Il est désormais possible de rechercher les photos en utilisant comme filtre une combinaison des étiquettes. Accepter l'imprécision inhérente à l'étiquetage, c'est accepter le désordre naturel du monde. C'est un antidote aux systèmes plus précis qui tentent d'imposer un ordre factice au tohu-bohu de la réalité, et professent que tout et n'importe quoi peut parfaitement s'intégrer dans des lignes et colonnes bien ordonnées. La chose est en réalité bien plus complexe qu'il n'y paraît.

De nombreux sites Web parmi les plus populaires affichent leur admiration pour l'imprécision plutôt que pour l'apparence d'exactitude. Quand on voit une icône correspondant à Twitter ou un bouton « J'aime » dans Facebook sur une page Web, ils indiquent le nombre d'autres personnes qui ont cliqué dessus. Quand il y en a peu, chaque clic est pris en compte et le nombre exact s'affiche, « 63 » par exemple. Mais au fur et à mesure que les chiffres augmentent, ce nombre est indiqué sous forme approximative, comme « 4 K », non parce que le système ne connaît pas le total exact, mais plutôt parce que l'indication des chiffres exacts a moins d'importance quand l'échelle augmente. En fait, les quantités peuvent changer si rapidement qu'un nombre spécifique serait obsolète à l'instant même de son affichage. De même, la messagerie Gmail de Google indique avec exactitude l'instant d'arrivée des messages récents, comme « il y a 11 minutes », mais pour les durées plus longues, elle formule un plus vague « il y a 2 heures », comme le font Facebook et certains autres outils.

Le secteur des logiciels de veille économique et des logiciels analytiques d'entreprise a longtemps promis à ses clients « une version unique de la vérité » – la formule à la mode dans les années 2000 des fournisseurs de ce genre de technologies. Les cadres employaient cette expression sans la moindre ironie – et certains d'ailleurs continuent de le

faire. Ils entendaient par là que toute personne ayant accès aux systèmes de technologie de l'information d'une entreprise pouvait exploiter les mêmes données ; que l'équipe marketing et la force de vente n'avaient pas besoin de se battre avant même le début de la réunion, pour déterminer qui détenait le nombre exact de clients ou les chiffres de ventes corrects. Leurs intérêts auraient été sans doute en meilleure harmonie si les faits étaient cohérents, comme on le pense généralement.

Mais aujourd'hui, on effectue un virage à 180 degrés à propos de cette idée d'« une version unique de la vérité ». Non seulement on se rend compte qu'il y a peu de chances qu'il y ait une seule version de la vérité, mais aussi qu'on fait fausse route en partant à sa recherche. Pour tirer les avantages de l'exploitation de données à grande échelle, il faut accepter que le désordre fasse partie du cours normal des choses, et qu'il ne s'agit pas d'un facteur à éliminer.

La culture de l'inexactitude commence même à envahir l'un des domaines les plus intolérants à l'imprécision : la conception de bases de données. Les moteurs de bases de données traditionnels exigeaient des données extrêmement structurées et précises. Les données n'étaient pas simplement stockées ; elles étaient subdivisées en « enregistrements » eux-mêmes subdivisés en champs. Chaque champ contenait des informations d'un type et d'une longueur spécifiques. Par exemple, pour un champ numérique d'une longueur de sept chiffres, il n'était pas possible d'enregistrer un montant de 10 millions ou plus. Si l'on voulait saisir « non disponible » dans un champ pour des numéros de téléphone, c'était impossible. Il aurait fallu modifier la structure de la base de données pour intégrer ces entrées. Nous continuons à nous heurter à ces restrictions imposées à nos ordinateurs et smartphones : le logiciel n'accepte pas certaines données que nous voulons y faire entrer.

Les index classiques eux aussi étaient prédéfinis, ce qui limitait les possibilités de recherches. Si l'on ajoutait un nouvel index, il fallait le créer de toutes pièces, et cela prenait du temps. Les bases de données classiques, appelées relationnelles, sont conçues pour un monde où les données sont peu abondantes, où il est donc possible de les organiser – un monde où il faut des questions claires dès le départ, auxquelles on veut trouver des réponses à partir des données, la base de données étant conçue de façon à y répondre efficacement.

Le problème est que cette conception du stockage et de l'analyse se retrouve de plus en plus en contradiction avec la réalité. Nous disposons maintenant de vastes quantités de données de type et de qualité variables. Et elles n'entrent que rarement dans des catégories nettement définies qu'on connaît dès le départ. De surcroît, c'est souvent après avoir collecté les données et quand nous nous en servons que les questions émergent.

Devant ce principe de réalité, il a fallu mettre sur pied des bases de données adoptant une conception innovante qui rompe avec les principes d'antan – enregistrements et champs préétablis correspondant exactement à des informations précisément hiérarchisées. L'accès aux bases de données a longtemps été régi par le langage SQL, le plus couramment utilisé, ou « langage de requête structuré » (le nom lui-même évoque sa rigidité). La grande révolution de ces dernières années a été de passer aux systèmes appelés noSQL, qui ne requièrent aucune structure d'enregistrement prédéfinie pour fonctionner. Le noSQL accepte les données de type et de taille très variés, ce qui permet une recherche efficace des données. En contrepartie de ce désordre structurel, les moyens de traitement et de stockage doivent être plus importants, mais un tel compromis est acceptable vu la chute des coûts.

Ce changement fondamental a été souligné par Pat Helland, l'une des autorités mondiales en matière de conception de bases de données, dans un article intitulé « Si vous avez trop de données, alors “assez bon” est vraiment assez bon ». Après avoir précisé que certains principes fondamentaux de la conception classique des bases de données ont été mis à mal par l'utilisation de données déstructurées de provenance et d'exactitude variables, il expose les conséquences : « Nous ne pouvons plus feindre de vivre dans un monde bien rangé. » Le traitement de volumes gigantesques de données entraîne une inévitable perte d'informations – Helland appelle ce phénomène « avec perte ». Il est compensé par l'obtention d'un résultat rapide. « Ce n'est pas un problème si nous avons des réponses avec perte – c'est souvent ce qu'attendent les entreprises », conclut Helland.

Une base de données de conception classique fournit des résultats constants au fil du temps. Si vous demandez le solde de votre compte bancaire, par exemple, vous vous attendez à obtenir le montant exact. Et si vous refaites la demande quelques secondes plus tard, vous voulez que le système vous fournisse le même résultat, en presumant que rien n'a changé. Mais il est de plus en plus difficile de conserver cette constance, avec

l'augmentation en volume du nombre de données ainsi que d'utilisateurs du système.

Les vastes ensembles de données ne sont pas localisés en un emplacement unique ; ils sont répartis dans de multiples disques durs et ordinateurs. Pour garantir fiabilité et rapidité, un enregistrement peut être stocké à deux ou trois emplacements distincts. Si vous mettez à jour l'enregistrement situé à un endroit, vous devez également mettre à jour les données placées ailleurs pour qu'elles soient correctes. Avec les systèmes classiques, il fallait un certain délai pour compléter toutes les mises à jour ; tout sauf pratique quand les données sont largement réparties et que le serveur est assailli de dizaines de requêtes par seconde. Dans ce cas, accepter le désordre est une forme de solution.

Hadoop, concurrent open source du système MapReduce de Google, très performant pour le traitement de grandes quantités de données, connaît un succès qui illustre bien ce changement. Il procède en décomposant les données en blocs plus petits et en les répartissant dans plusieurs machines. Comme il prévoit les défaillances possibles du matériel informatique, il intègre donc une dimension de fonctionnement redondant. Il fait la supposition que les données ne sont ni apurées ni ordonnées – en fait, il prend pour acquis qu'elles ne seront pas nettoyées avant traitement car trop volumineuses. Là où l'analyse classique des données commence par l'opération d'« extraction, transfert et chargement » (ETL) avant de déplacer les données vers l'emplacement où elles seront analysées, Hadoop se dispense de ces subtilités. La quantité de données est telle, estime Hadoop, qu'il n'est pas question de la déplacer et il faut l'analyser sur place.

Le résultat n'est pas aussi précis que celui des bases de données relationnelles classiques : il ne faudra pas s'en servir pour lancer un vaisseau spatial ou certifier les coordonnées d'un compte bancaire. Mais pour bon nombre de tâches moins essentielles, qui n'ont pas besoin d'une réponse ultra-précise, Hadoop trouve une solution bien plus rapidement que les autres systèmes – entre autres tâches, segmenter une liste de consommateurs pour adresser à certains d'entre eux une campagne marketing spécifique. Grâce à Hadoop, le temps de traitement d'enregistrements de la société de cartes de crédit Visa, correspondant à un volume de près de 73 milliards de transactions sur deux ans, a pu être réduit

à 13 minutes à peine au lieu d'un mois. Une telle accélération du traitement transforme les entreprises.

À preuve, l'expérience de la société ZestFinance, créée par Douglas Merrill, ex-numéro un de l'information chez Google. Sa technologie permet à des prêteurs de décider s'ils proposent ou pas des prêts à court terme d'un montant relativement faible à des personnes qui semblent de faible solvabilité. Là où l'enquête de solvabilité classique ne se fonde que sur quelques indices forts comme des retards de paiement antérieurs, ZestFinance analyse un certain nombre de variables plus « faibles ». En 2012, les défauts de remboursement de prêt affichaient un taux inférieur d'un tiers par rapport à la moyenne du secteur. Pour que ce système fonctionne, la seule solution était d'y intégrer le désordre.

« Il est intéressant de remarquer, déclare Merrill, que personne ne remplit absolument tous les champs – il existe toujours une bonne quantité de données manquantes. » Le tableau créé à partir des informations collectées par ZestFinance est incroyablement clairsemé, un fichier où les cases vides abondent. Aussi la société « attribue »-t-elle les données manquantes. Par exemple, près de 10 % des clients de ZestFinance figurent dans la liste comme décédés – ce qui, en fait, n'a pas l'air d'affecter le remboursement. « Évidemment, quand ils se préparent à l'apocalypse des zombies, la plupart des gens supposent qu'aucune dette ne sera remboursée. Mais d'après nos données, il apparaît que les zombies remboursent leurs prêts », ajoute Merrill avec un clin d'œil.

En contrepartie d'une cohabitation avec le désordre, nous obtenons des services extrêmement précieux, chose impossible à obtenir à pareille échelle avec des méthodes et des outils classiques. D'après certaines estimations, 5 % seulement de toutes les données numériques sont « structurées » – c'est-à-dire sous une forme parfaitement adaptée à une base de données classique. À moins d'accepter le désordre, les 95 % restants, comme les pages Web et les vidéos, demeurent opaques. Une fois l'imprécision acceptée, c'est un univers inexploité d'informations qui s'ouvre.

La société a fait deux compromis implicites que nous avons fini par si bien intégrer que nous ne les voyons même plus ; pour nous, ils vont de soi. En premier, nous ne tentons pas d'utiliser beaucoup plus de données car nous partons du principe que nous ne pouvons pas le faire. Mais cette

contrainte se justifie de moins en moins, et il y a tout intérêt à utiliser un nombre de données proche de la formule $N = \text{tous}$.

Le second compromis porte sur la qualité des informations. Il était rationnel de privilégier l'exactitude à l'époque où l'on utilisait de faibles volumes de données. Dans de nombreux cas, cela peut encore avoir une importance. Mais autrement, il est moins important de détenir des informations rigoureusement exactes que de saisir rapidement leurs grandes lignes ou leur évolution avec le temps.

Il faut s'attendre à de grands changements dans notre interaction avec le monde, si l'on suit l'idée d'utiliser la totalité plutôt que des bouts d'informations et de raisonner en termes d'imprécision et de probabilités plutôt que d'exactitude. Maintenant que les techniques de big data commencent à faire partie intégrante de notre vie quotidienne, nous pourrions en tant que société commencer à tenter d'appréhender le monde avec une vision bien plus vaste et plus globale qu'auparavant, une sorte de formule $N = \text{tous}$ de l'esprit. Et tolérer un flou et une ambiguïté dans des domaines où nous exigeons autrefois de la clarté et une certitude, quitte à ce que cette clarté soit fausse et cette certitude imparfaite. Ce serait acceptable à condition qu'en contrepartie on puisse bénéficier d'un sens plus complet de la réalité – l'équivalent d'une peinture impressionniste, où chaque coup de pinceau est déstructuré quand on l'observe de près, mais dont l'ensemble révèle un tableau majestueux lorsqu'on prend du recul pour le regarder.

Les big data, autrement dit des ensembles immenses de données déstructurées, nous rapprochent plus de la réalité que ne le faisaient des volumes plus petits de données aussi exactes que possible. L'attrait de notions telles que « quelques » et « certaines » est compréhensible. Nous avions une appréhension du monde sans doute incomplète et parfois erronée quand le volume de données à analyser demeurait à un niveau limité, mais c'était une situation confortable, à la stabilité rassurante. De surcroît, parce que nous ne pouvions collecter et étudier qu'une quantité limitée de données, nous n'étions pas confrontés à ce besoin irrésistible de tout voir sous tous les angles possibles et de tout avoir. Coincés dans des limites étroites, nous nous félicitions de notre précision – même si, à force de mesurer les moindres détails, l'ensemble du tableau nous échappait.

In fine, il se pourrait qu'avec les big data, nous devions changer, afin de nous sentir plus à l'aise avec les notions de désordre et d'incertitude. Il y a plus de souplesse dans les structures que la seule exactitude qui semble

nous servir de repère : le bloc rond va dans le trou rond ; il n'y a qu'une seule réponse à chaque question. Il nous faudra admettre cette notion de plasticité, et même l'adopter dans la pratique, car cela nous rapproche plus de la réalité.

Ces modifications de notre état d'esprit, outre la transformation radicale qu'elles entraînent, augurent d'un troisième changement fondamental pour la société. Alors que nous voulons systématiquement comprendre les raisons qui sous-tendent tout ce qu'il advient dans le monde, nous pouvons souvent nous contenter des corrélations qu'il est possible de découvrir au sein des très grands ensembles de données. C'est ce que nous allons expliquer au chapitre suivant.

- [a.](#) The Brown University Standard Corpus of Present-Day American English : corpus de l'anglais américain actuel.
- [b.](#) L'efficacité déraisonnable des données.
- [c.](#) Bureau américain des statistiques sur le travail.

4

Corrélation

En 1997, Greg Linden, à l'âge de vingt-quatre ans, alors qu'il menait des recherches en intelligence artificielle dans le cadre d'un doctorat à l'université de Washington, décida de prendre du temps sur ses études pour travailler dans une start-up Internet locale qui vendait des livres en ligne. Créée seulement deux ans auparavant, elle était déjà florissante. « J'aimais l'idée de vendre des livres et du savoir – et d'aider les gens à trouver un nouvel ouvrage qui répondrait à leurs attentes et enrichirait leurs connaissances », se souvient-il. Ce magasin était Amazon.com, et Linden y fut recruté au titre d'ingénieur informatique chargé de veiller au bon fonctionnement du site.

Le personnel d'Amazon ne se composait pas exclusivement de techniciens. À l'époque, la société employait également près d'une dizaine de critiques littéraires et de rédacteurs dont la tâche consistait à écrire des comptes rendus et à suggérer de nouveaux titres. Si l'histoire d'Amazon est bien connue, peu de gens se souviennent qu'à l'origine, son contenu était « fait main ». Les rédacteurs et les critiques qui évaluaient et choisissaient les titres figurant sur les pages Web d'Amazon avaient pour mission de faire entendre ce qui s'appelait la « voix d'Amazon » – considérée comme l'un des « joyaux de la couronne » de l'entreprise et source de son avantage concurrentiel. Selon un article du *Wall Street Journal* de l'époque, ils s'étaient élevés au rang de critiques littéraires les plus influents du pays, en raison du rôle qu'ils jouaient dans la formidable croissance des ventes.

Puis Jeff Bezos, fondateur et président-directeur général d'Amazon, commença à expérimenter une idée forte : et si la société arrivait à recommander des ouvrages spécifiques en fonction des goûts personnels des clients ? Dès sa création, Amazon avait recueilli une flopée de données

sur ces derniers : la trace de leurs achats, mais aussi des livres qu'ils se contentaient de regarder sans acheter et le temps que ça leur avait pris, les livres qu'ils achetaient en une seule commande...

La quantité de données était si énorme qu'au début, Amazon les traita selon la méthode classique, en prenant un échantillon et en l'analysant pour trouver des similarités entre clients. Les recommandations qui s'ensuivirent étaient extrêmement simplistes. Après l'achat d'un livre sur la Pologne, vous étiez bombardé de toute la littérature disponible sur l'Europe de l'Est. Idem s'il s'agissait d'un livre sur les bébés, vous étiez inondé de titres afférents. « La tendance était de vous proposer de légères variations sur le thème de votre précédent achat, *ad infinitum*, se rappelle le critique littéraire pour Amazon de 1996 à 2001, James Marcus, dans ses Mémoires, *Amazonia*. Cela donnait l'impression d'aller faire ses emplettes en compagnie de l'idiot du village. »

C'est alors que Greg Linden a entrevu une solution. Pour émettre des recommandations, il n'était pas besoin que le système compare des personnes à d'autres, opération techniquement lourde à gérer. Il suffisait de trouver des associations entre les produits eux-mêmes. En 1998, Linden et ses collègues déposèrent une demande de brevet pour une technique de filtrage collaboratif de ce qu'ils appelèrent « article à article ». Avec ce changement de méthode, la différence fut très nette.

Du fait que les calculs pouvaient être effectués à l'avance, les recommandations ont pu s'afficher à la vitesse de l'éclair. La méthode polyvalente a pu être adaptée à des produits de diverses catégories. Si bien que quand Amazon s'est diversifié pour vendre d'autres articles outre les livres, le système a pu tout aussi bien suggérer des films que des grille-pain. L'utilisation de toutes les données a permis de suggérer des recommandations bien meilleures qu'auparavant. « La plaisanterie qui courait dans le groupe était que si Amazon atteignait la perfection, elle pourrait simplement vous indiquer un seul et unique livre – votre prochain achat », se souvient Linden.

Mais la société a dû alors faire un choix : le site allait-il présenter un contenu généré par des machines (recommandations personnelles et listes des meilleures ventes) ou bien des comptes rendus rédigés par l'équipe éditoriale d'Amazon ? Fallait-il se fonder sur ce que les clics révélaient ou sur ce que les critiques disaient ? L'heure était venue de la confrontation hommes-machines.

Quand Amazon a fait un test de comparaison entre les ventes réalisées grâce aux éditeurs humains et celles dues au contenu généré par l'ordinateur, les résultats furent sans appel : le second engendrait un nombre bien supérieur de ventes que les premiers. Sans aucun doute, l'ordinateur ignorait pourquoi un client lecteur des œuvres d'Ernest Hemingway aurait également envie d'acheter du F. Scott Fitzgerald. Aucune importance, le chiffre d'affaires augmentait. Amazon finit par présenter aux éditeurs le pourcentage précis de ventes ratées lors de la publication de leurs comptes rendus sur le site et le groupe fut donc dissous. « J'étais très triste de l'échec rencontré par l'équipe éditoriale, se rappelle Linden. Mais les données ne mentent pas, et cela coûtait très cher à la société. »

Actuellement, un tiers de toutes les ventes d'Amazon seraient imputables à ses systèmes de recommandation et de personnalisation. Et dans la foulée, Amazon a précipité la faillite de nombreux concurrents : pas seulement de grandes librairies et de magasins de musique, mais aussi des librairies de quartier qui pensaient que leur touche personnelle les protégerait des vents du changement. En fait, le travail de Linden a révolutionné le commerce électronique, et sa méthode a été adoptée par quasiment tout le monde. Pour Netflix, société de location de films en ligne, trois quarts des nouvelles commandes viennent des recommandations. Dans le sillage d'Amazon, des milliers de sites Web recommandent maintenant des produits, des contenus, des amis et des groupes sans savoir pourquoi les gens sont susceptibles de s'y intéresser.

En connaître la raison réserverait sans doute d'agréables surprises, mais ne présenterait aucun intérêt pour stimuler les ventes. En revanche, il faut savoir quoi recommander pour générer des clics. Cette connaissance a le pouvoir de transformer de nombreux secteurs, et pas seulement le commerce électronique. On a expliqué pendant longtemps aux commerciaux dans tous les secteurs qu'ils devaient comprendre ce qui déclenchait la commande du client, saisir les raisons sous-tendant ses décisions. On a mis en exergue les compétences professionnelles et les années d'expérience. Les big data démontrent qu'il existe une autre démarche, plus pragmatique à certains égards. Les systèmes innovants de recommandation mis au point par Amazon ont fait ressortir des corrélations efficaces sans en connaître les causes sous-jacentes. Savoir *quoi*, et non *pourquoi*, est largement suffisant.

Prédictions et prédilections

Dans un monde de petits volumes de données, les corrélations sont utiles mais dans le contexte des big data, elles font merveille. Grâce à elles, nous obtenons des informations plus facilement, plus rapidement et de façon plus claire qu'auparavant.

Une corrélation quantifie la relation statistique entre deux valeurs. Elle est forte si une valeur a de fortes chances de changer quand la première est modifiée. Nous avons observé des corrélations fortes de ce type chez Google Suivi de la grippe : plus il y a de gens d'une localité particulière qui recherchent des termes spécifiques sur Google, plus il y a de résidents atteints de la grippe. À l'inverse, la corrélation est faible si une valeur varie peu alors que la première a changé. Par exemple, nous pourrions établir des corrélations entre la longueur des cheveux et le bonheur et découvrir que la première ne renseigne pas beaucoup sur le second.

L'analyse d'un phénomène par des corrélations ne nous renseigne pas sur son fonctionnement interne mais nous permet d'identifier une donnée utile qui s'y substitue. Bien sûr, même quand elles sont fortes, les corrélations ne sont jamais parfaites. Il est tout à fait possible que deux choses se comportent de manière similaire par pure coïncidence. Nous pouvons être tout simplement « trompés par le hasard », pour reprendre l'expression de l'empiriste Nassim Nicholas Taleb^a. Avec les corrélations, il n'y a aucune certitude, uniquement de la probabilité. Mais si une corrélation est forte, cela augmente la probabilité qu'il existe un lien. De nombreux clients d'Amazon peuvent en témoigner en montrant une étagère remplie de livres acquis suite aux recommandations de l'entreprise.

En nous permettant d'identifier une très bonne donnée substitutive pour un phénomène, les corrélations nous aident à saisir le présent et à prédire l'avenir : si A et B ont souvent lieu en même temps, nous devons guetter l'apparition de B pour prédire que A se produira. Utiliser B comme donnée substitutive nous permet de saisir ce qui se passe probablement pour A, même si nous ne pouvons mesurer ou observer A directement. Chose importante, cela nous aide également à prédire ce qui pourrait arriver à A dans le futur. Naturellement, les corrélations ne peuvent prédire l'avenir de façon infaillible, elles ne peuvent le prédire qu'en tenant compte d'une certaine probabilité. Mais cette capacité est extrêmement précieuse.

Prenons le cas de Walmart, le plus grand distributeur mondial qui compte plus de deux millions d'employés et affiche un chiffre d'affaires de

près de 450 milliards de dollars – montant supérieur au PIB de quatre pays sur cinq dans le monde. Avant que le Web ne produise autant de données, la société était, parmi toutes les entreprises américaines, celle qui détenait sans doute le plus grand ensemble de données. Dans les années 1990, elle a révolutionné la distribution en décidant que chaque produit serait enregistré sous forme de donnée grâce à un système appelé « Retail Link », ce qui a permis à ses fournisseurs de contrôler le débit et le volume des ventes ainsi que l'inventaire. Grâce à ce suivi parfaitement transparent, la société a obligé ses fournisseurs à s'occuper eux-mêmes du stockage. Dans de nombreux cas, la non-prise en charge du produit par Walmart jusqu'au point de vente lui permet de se désengager du risque lié au stockage et de réduire ses coûts. Walmart s'est servi de données pour devenir, de fait, le plus grand dépôt-vente du monde.

Que pourraient révéler ces données historiques si elles étaient analysées de façon adéquate ? Que des corrélations intéressantes existent, comme le distributeur l'a découvert en collaborant avec des experts en calculs de Teradata, l'ancienne et vénérable National Cash Register Company. En 2004, en fouillant dans ses gigantesques bases de données de transactions passées, Walmart a recherché la trace de l'achat de tel ou tel article par tel ou tel client, a retrouvé le coût total de la transaction, ainsi que ce qu'il y avait d'autre dans le panier d'achats, ou encore le moment de la transaction dans la journée, et même le climat ce jour-là. Grâce à ces recherches, la société a pu s'apercevoir qu'avant un ouragan, il y avait plus de ventes de lampes de poche, mais également une augmentation des ventes de Pop-Tarts, biscuits fourrés américains. Ainsi, à l'approche de tempêtes, Walmart s'est mis à présenter des paquets de Pop-Tarts à l'entrée des magasins juste à côté des articles utiles aux clients lors d'ouragans et a vu ses ventes grimper.

Jadis, il aurait fallu l'intuition d'une personne du siège pour que pareille initiative soit lancée puis que les données soient recueillies et mesurées. Maintenant, avec les quantités de données et les meilleurs outils d'analyse dont on dispose, les corrélations émergent plus rapidement et à moindre coût. (Cela dit, il faut rester prudent : quand le nombre de points de données augmente de plusieurs ordres de grandeur, les corrélations trompeuses aussi – des phénomènes apparaissent liés même s'ils ne le sont pas vraiment. Il faut donc redoubler d'attention, étant donné que nous commençons à peine à en prendre conscience.)

Bien avant les big data, l'utilité de l'analyse des corrélations avait été démontrée. Le concept a été formulé en 1888 par Sir Francis Galton, cousin de Charles Darwin, qui avait noté chez les humains une relation entre leur taille et la longueur des avant-bras. Cette observation est sous-tendue par un calcul mathématique relativement simple et fiable – c'est d'ailleurs l'un de ses aspects essentiels qui a permis d'en faire l'une des mesures statistiques les plus largement utilisées. Avant les big data, son utilité était limitée. Les données étaient rares, leur collecte coûteuse, et par conséquent les statisticiens choisissaient souvent une donnée substitutive, puis collectaient les données pertinentes et effectuaient l'analyse des corrélations pour connaître le degré d'efficacité de cette donnée substitutive. Mais comment choisir la donnée substitutive adéquate ?

Pour les guider dans leur choix, les experts ont formulé des hypothèses théoriques – des idées abstraites sur la façon dont quelque chose fonctionne. En partant de ces hypothèses, ils ont collecté des données et l'analyse de corrélations leur a permis de vérifier si les données substitutives convenaient. Dans le cas contraire, les chercheurs ont souvent dû procéder à de nouveaux essais, en s'assurant que la procédure de collecte des données ne soit pas erronée, avant de finir par concéder que leur hypothèse de départ, ou même la théorie sur laquelle elle était fondée, était incorrecte et nécessitait d'être modifiée. Les connaissances ont progressé par tâtonnements à partir d'une hypothèse, mais lentement, étant donné que nos a priori individuels et collectifs ont semé la confusion lors de l'élaboration des hypothèses, lors de leur application, et par conséquent lors du choix de données substitutives. La procédure a été laborieuse, mais praticable dans un monde de petits volumes de données.

À l'ère des big data, prendre des décisions en se fondant sur l'examen des variables et en s'appuyant uniquement sur des hypothèses a perdu de son efficacité. La taille des ensembles de données est bien trop grande et la complexité du domaine étudié probablement bien trop forte. Heureusement, les nombreuses limites qui nous obligeaient à opter pour une méthode fondée sur l'hypothèse n'ont plus la même importance. Grâce aux très nombreuses données et à une grande puissance de calcul, nous ne sommes pas contraints de choisir laborieusement une ou plusieurs donnée(s) substitutive(s) puis de les examiner une à une. On peut aujourd'hui identifier la donnée substitutive optimale grâce à une analyse élaborée,

comme ce fut le cas pour Google Suivi de la grippe, après qu'il eut passé en revue près d'un demi-milliard de modèles mathématiques.

Pour commencer à comprendre le monde où nous vivons, il n'est plus nécessaire de disposer à tout prix d'une hypothèse cohérente sur un phénomène. Point n'est besoin d'imaginer les termes utilisés dans les requêtes des gens si nous voulons rechercher des informations sur le moment et le lieu de la propagation de la grippe. Nous n'avons pas besoin d'avoir une idée sur la méthode employée par les compagnies d'aviation pour fixer les prix de leurs billets ni de nous préoccuper des goûts culinaires des clients de Walmart. Il nous suffit de soumettre les big data à une analyse des corrélations de façon à repérer quelles requêtes sont les meilleures données substitutives pour la grippe, si le prix d'un billet d'avion est susceptible de monter en flèche, ou bien combien de familles angoissées veulent grignoter des douceurs durant une tempête. En lieu et place de la méthode fondée sur des hypothèses, nous utilisons celle qui se base sur les données. Nos résultats seront probablement plus objectifs et plus exacts, et nous les obtiendrons (c'est quasi certain) bien plus rapidement.

Les prédictions basées sur des corrélations sont au cœur des big data. Les analyses de corrélations sont maintenant utilisées si fréquemment que, parfois, nous n'estimons pas à sa juste mesure à quel point elles permettent une percée, ni à quelle hauteur l'utilisation qui peut en être faite va se multiplier.

Par exemple, les pointages de crédit. Ils sont actuellement utilisés pour prédire le comportement personnel. La Fair Isaac Corporation, maintenant connue sous le nom de FICO, a inventé les pointages de crédit à la fin des années 1950. En 2011, FICO a créé un score intitulé « Score de respect du traitement médical ». Dans le but de déterminer la probabilité de suivi de leur traitement que présentent les gens, FICO analyse une multitude de variables, dont certaines peuvent paraître dénuées de pertinence : depuis combien de temps ces gens vivent à la même adresse, s'ils sont mariés, depuis combien de temps ils occupent le même poste et s'ils possèdent un véhicule. Le score est destiné à indiquer aux prestataires de santé sur quels patients cibler leurs rappels tout en économisant de l'argent. Il n'y a pas de relation de cause à effet entre la possession d'un véhicule et le respect de la prescription d'antibiotiques ; le lien entre eux est une pure corrélation. Mais c'est sur la base de conclusions de ce type que le président-directeur général

de FICO, en 2011, s'est vanté d'affirmer : « Nous savons ce que vous ferez demain. »

La série originale du *Wall Street Journal* intitulée « Ce qu'ils savent » a montré que d'autres courtiers en données se lancent eux aussi sur le terrain des corrélations. Grâce à son produit appelé Income Insight, la société Experian estime le niveau de revenu des gens en se fondant en partie sur leurs antécédents de crédit. C'est en analysant son énorme base de données des antécédents de crédit et en les comparant aux données anonymes sur les impôts, issues de l'U.S. Internal Revenue Service^b, qu'elle y est parvenue. Alors qu'il en coûterait à une société près de 10 dollars pour qu'elle ait la confirmation du revenu d'une personne à partir des formulaires fiscaux, Experian vend son estimation à moins d'un dollar. Dans ce type de cas, utiliser la donnée substitutive est donc plus rentable que de passer par tout le scénario d'obtention de la donnée réelle elle-même. De la même manière, un « Indice de solvabilité » et un « Indice des dépenses discrétionnaires », qui prédisent l'épaisseur du portefeuille d'une personne, sont vendus par un autre organisme de crédit, Equifax, qui assure de son sérieux.

Les corrélations vont même plus loin dans leur utilisation. À la place des résultats d'analyse d'échantillons de sang et d'urine qu'il faut parfois produire lors d'une demande d'assurance, la grande société d'assurance Aviva a imaginé d'utiliser comme données substitutives les rapports de crédit et les données commerciales relatives aux consommateurs. Objectif : identifier ceux qui présentent un risque plus grand de pathologies comme l'hypertension artérielle, le diabète ou la dépression. La méthode ? Utiliser les données concernant le style de vie et comprenant des centaines de variables relatives aux loisirs, à la visite de tel ou tel site Web, au nombre d'heures passées devant la télévision ainsi que l'estimation du revenu.

Le modèle prédictif d'Aviva, développé par Deloitte Consulting, a été jugé efficace pour identifier les risques de santé. D'autres compagnies d'assurances comme Prudential et AIG ont envisagé des initiatives du même genre qui ont pour avantage d'éviter, lors d'une demande de couverture d'assurance, d'avoir à fournir des échantillons de sang et d'urine, analyse qui n'est agréable pour personne et dont le coût doit être acquitté par les compagnies d'assurances. Il est d'environ 125 dollars par personne, contre 5 dollars environ pour la méthode reposant uniquement sur des données.

La méthode peut paraître bizarre à certains, puisqu'elle se fonde sur des comportements apparemment sans lien. C'est comme si des sociétés se dotaient d'un cyber-mouchard qui espionnerait chaque clic de souris. Les gens y réfléchiraient alors à deux fois avant de visiter des sites Web de sports extrêmes ou de regarder des sitcoms glorifiant les téléphages, s'ils pensaient que cela pouvait entraîner pour eux une hausse de leur prime d'assurance. Une restriction de la liberté individuelle qui consiste à chercher et recevoir des informations serait certes fâcheuse. D'un autre côté, cela présente un avantage : avec une couverture d'assurance moins chère et plus facile à obtenir, le nombre d'assurés pourrait augmenter, ce qui est une bonne chose pour la société, et a fortiori pour les compagnies d'assurances.

L'utilisateur à la fois typique, et victime si l'on peut dire, des corrélations des big data est celui qui se sert chez le détaillant discount américain Target. Voilà des années que ce dernier fait appel aux corrélations obtenues à partir de big data pour prédire ses ventes. Dans un extraordinaire article, Charles Duhigg, correspondant économique au *New York Times*, a montré comment Target est au courant de la grossesse d'une femme sans que celle-ci le dise explicitement. À la base, sa méthode consiste à exploiter des données et laisser les corrélations faire leur « boulot ».

Savoir si une cliente attend un enfant est important pour les détaillants, étant donné que la grossesse joue un rôle décisif dans la vie des couples et est susceptible de modifier leur comportement d'achat. Ils pourraient fréquenter de nouveaux magasins et être fidélisés par de nouvelles marques. Chez Target, les spécialistes en marketing se sont adressés à leur département analytique pour voir s'il y avait moyen de découvrir la grossesse des clientes à travers leurs habitudes d'achat.

L'équipe du département a étudié les historiques d'achats de celles qui avaient créé une liste de naissance. Ils ont remarqué que ces femmes achetaient beaucoup de lotion non parfumée vers le troisième mois de grossesse et qu'elles avaient tendance, quelques semaines plus tard, à rechercher des compléments alimentaires – magnésium, calcium et zinc... En fin d'étude, l'équipe avait mis au jour près d'une vingtaine de produits qui pouvaient être utilisés comme données substitutives. Ce qui a permis à la société de calculer un score de « prédiction grossesse » pour chacune des clientes réglant leurs achats avec une carte de crédit ou bien utilisant une carte de fidélité ou des bons d'achat reçus par la poste. Grâce aux

corrélations, il a été possible au détaillant d'estimer la date de l'accouchement dans une étroite fourchette, ce qui a permis en retour de faire parvenir des bons d'achat correspondant à chaque étape de la grossesse. La société porte bien son nom de « Cible^c » !

Dans son livre *Le Pouvoir des habitudes*^d, Duhigg raconte ce qui s'est passé ultérieurement. Un jour, un homme est entré furieux dans un magasin Target du Minnesota pour y voir un directeur. « Ma fille a reçu ce courrier ! hurla-t-il. Elle va encore au lycée, et vous lui envoyez des bons d'achat pour des vêtements de bébé et des berceaux ? Vous essayez de l'encourager à tomber enceinte ? » Mais quand le directeur appela le monsieur quelques jours plus tard pour lui présenter des excuses, sa voix était devenue conciliante. « J'ai eu une discussion avec ma fille, dit-il. Il s'avère qu'il s'est passé des choses sous mon toit dont je ne m'étais pas vraiment rendu compte. La naissance est prévue pour août. C'est à moi de vous présenter des excuses. »

Trouver des données substitutives dans divers contextes sociaux n'est qu'une des manières d'utiliser ces techniques de big data. Répondre à des besoins de tous les jours au moyen de corrélations avec de nouveaux types de données se révèle tout aussi efficace.

Il existe par exemple une méthode baptisée analyse prédictive, qui commence à être largement utilisée dans le monde du commerce pour prévoir les événements avant qu'ils n'arrivent. Le terme peut désigner un algorithme capable de repérer un futur « tube », couramment utilisé dans l'industrie musicale pour donner aux marques de disques une meilleure idée des titres sur lesquels miser. La technique est également utilisée pour prévenir de graves défaillances mécaniques ou structurelles : on peut contrôler des machines, des moteurs ou des infrastructures comme les ponts, en y plaçant des capteurs qui recueillent de nombreux types de données – chaleur, vibration, contraintes et sons, ce qui permet de détecter des changements annonciateurs de problèmes à venir.

Cette méthode se fonde sur le concept selon lequel une défaillance, quand elle survient, ne se produit en général pas d'un seul coup, mais graduellement. À l'aide de capteurs qui enregistrent les données, puis grâce à l'analyse par corrélation ou autres méthodes similaires, il est possible d'identifier les signaux avant-coureurs d'une panne – vrombissement d'un moteur, température excessive d'un moteur thermique, etc. À partir de là, il n'y a plus qu'à repérer ce signal pour savoir quand quelque chose ne va pas.

Détecter précocement l'anomalie permet au système d'envoyer un avertissement pour qu'il soit procédé à l'installation d'une nouvelle pièce ou que le problème soit réglé avant que ne survienne effectivement la panne. L'objectif est de trouver la bonne donnée substitutive puis de la guetter, et ainsi de prédire des incidents futurs.

Depuis la fin des années 2000, la société de transport express de colis UPS contrôle son parc de 60 000 véhicules aux États-Unis au moyen de l'analyse prédictive, afin de savoir où effectuer une maintenance préventive. Une panne peut provoquer de graves perturbations et retarder livraisons et ramassages de colis. Jadis, par précaution, UPS remplaçait des pièces au bout de deux ou trois ans. Mais l'opération manquait d'efficacité car certaines de ces pièces étaient encore en bon état. Depuis qu'elle a adopté l'analyse prédictive et remplace les pièces uniquement lorsque cela se révèle nécessaire après mesure et contrôle, la société a économisé des millions de dollars. Il est même arrivé que les données aient révélé la présence d'une pièce défectueuse dans tout un groupe de nouveaux véhicules, qui aurait pu causer de sérieux problèmes si elle n'avait pas été repérée avant leur entrée en service.

Des capteurs sont également fixés aux ponts et aux bâtiments de façon à guetter les signes d'usure. Ils sont utilisés dans de grandes usines chimiques et des raffineries, où du matériel défectueux peut paralyser la production. La collecte et l'analyse des données qui indiquent à quel moment intervenir en amont coûtent moins cher qu'une panne. Il est à noter que l'analyse prédictive n'explique pas la cause d'un problème ; elle ne fait qu'indiquer son existence. Elle vous prévient qu'un moteur est en surchauffe, mais ne vous indiquera pas forcément si cette surchauffe est due à une courroie usée ou à un bouchon mal vissé. Avec les corrélations, on voit ce qui se produit et non la cause, mais comme nous l'avons déjà indiqué, cela suffit souvent.

Le même type de méthodologie est appliqué dans le domaine de la santé, pour prévenir les pannes de la machine humaine. Quand un patient hospitalisé est relié à un ensemble de tubes, de câbles et d'instruments, c'est un vaste flux de données qui est généré. À lui seul, l'électrocardiogramme enregistre 1 000 valeurs par seconde. Et pourtant, fait remarquable, seule une fraction des données est actuellement utilisée ou conservée. La plupart sont simplement jetées, même si elles recèlent des indications importantes sur l'état de santé du patient et comment il réagit aux traitements. Seraient-elles conservées et agrégées à d'autres données sur le patient, qu'elles

pourraient révéler d'extraordinaires informations sur la potentielle réussite (ou pas) de ces traitements.

On peut comprendre qu'on éliminait jadis des données quand le coût et la complexité de la collecte, du stockage et de l'analyse étaient élevés. Mais ce n'est plus le cas. Le docteur Carolyn McGregor avec une équipe de chercheurs à l'Institut de technologie de l'université de l'Ontario, la société IBM et un certain nombre d'hôpitaux ont ainsi décidé de collaborer à la mise au point d'un logiciel d'aide au diagnostic, permettant aux médecins de mieux prendre en charge des bébés prématurés. Le logiciel qui capture et traite en temps réel les données relatives au patient recueille 16 flux de données différents, tels les battements de cœur, le rythme de la respiration, la tension artérielle et le taux d'oxygène dans le sang. En tout, près de 1 260 points de données sont enregistrés chaque seconde.

D'infimes changements dans l'état de santé des prématurés peuvent être détectés par le système, qui annoncent le début d'une infection, 24 heures avant l'apparition des symptômes. « C'est impossible à voir à l'œil nu, mais un ordinateur peut la déceler », explique le docteur McGregor. Aucune causalité dans le système, mais des corrélations. Il indique les faits, non les raisons. Mais le but est atteint : être averti précocement permet au médecin de traiter l'infection plus tôt et de recourir à des interventions médicales plus légères, ou alors l'inefficacité d'un traitement est signalée plus rapidement. Ainsi, l'état de santé du patient s'améliore. Il y a de bonnes raisons de penser que dans l'avenir, cette technique sera appliquée à un plus grand nombre de patients et de pathologies. Certes, ce n'est pas l'algorithme lui-même qui prend les décisions, mais les machines font ce qu'elles savent faire le mieux, à savoir contribuer à l'optimisation des soins médicaux que prodiguent les intervenants humains.

Fait frappant, les corrélations révélées par l'analyse des big data mise au point par le docteur McGregor vont, par certains aspects, à l'encontre des idées reçues courantes chez les médecins. Elle a constaté par exemple qu'avant qu'une infection grave n'apparaisse, ce sont des signaux très constants qui sont souvent détectés par le système. C'est étonnant, car on aurait tendance à penser qu'une infection déclarée serait précédée par une détérioration des paramètres vitaux. Il est facile d'imaginer bien des générations de médecins jetant en fin de journée un coup d'œil à la pancarte placée près du berceau et constatant une stabilisation des signes vitaux du nourrisson, se dire qu'ils pouvaient rentrer l'esprit tranquille à la maison –

pour recevoir à minuit un appel téléphonique paniqué des infirmiers (ières) leur annonçant un événement tragique, leur instinct les ayant donc trompés.

D'après les données de McGregor, il semble que la stabilité chez le bébé soit moins un signe d'amélioration que d'accalmie avant la tempête – comme si son corps disait à ses minuscules organes de fermer les écoutilles avant une difficile traversée. Ce n'est pas une certitude : ce que les données indiquent est une corrélation, et non une causalité. Mais nous sommes sûrs d'une chose : cette association cachée a été révélée grâce à des méthodes statistiques appliquées à une énorme quantité de données. Qu'il n'y ait aucun doute à ce sujet : les big data sauvent des vies humaines.

Illusions et illuminations

Dans un monde de données en petites quantités, vu le faible volume disponible, la recherche des causes et l'analyse des corrélations commencent par la formulation d'une hypothèse de départ, qu'il faut ensuite tester afin de l'infirmier ou de la confirmer. D'où un risque de biais, dû à des idées préconçues et à une intuition erronée. Souvent, les données nécessaires ne se révèlent pas disponibles. Aujourd'hui, avec le grand nombre de données existantes et à venir, les hypothèses ne sont plus indispensables pour mener l'analyse par corrélation.

Autre différence qui prend de plus en plus d'importance : avant les big data, et en partie à cause d'une puissance de calcul inadéquate, une bonne partie de l'analyse se limitait à une recherche des relations linéaires. Évidemment, bon nombre de relations sont en réalité bien plus complexes. Grâce à des analyses plus élaborées, nous pouvons identifier des relations non linéaires entre des données.

Prenons, à titre d'exemple, l'existence d'une corrélation directe entre le bonheur et les revenus que, pendant des années, les économistes et les politologues ont cru pouvoir affirmer : que les revenus d'une personne augmentent et elle sera en règle générale plus heureuse. Mais si on regarde le graphique des données, on découvre qu'une dynamique plus complexe entre en jeu. Pour des revenus au-dessous d'un certain niveau, toute hausse se traduit par un sentiment de bonheur accru. Mais au-delà de ce seuil, ce sentiment devient à peine sensible aux augmentations de revenu. Si on en faisait une représentation graphique, on tracerait une courbe plutôt qu'une ligne droite telle que l'analyse linéaire le supposerait.

Un tel constat est d'importance pour les décideurs politiques. En cas de relation linéaire, il serait logique de rendre les citoyens plus heureux en augmentant leur revenu. Sauf que la découverte de l'association non linéaire a modifié la recommandation : c'est l'augmentation de revenu des pauvres qui est la plus profitable, montrent les données.

Et les relations peuvent devenir encore plus complexes : par exemple quand les corrélations comportent plusieurs facettes. Il en va ainsi de la disparité en matière de vaccination contre la rougeole, comme l'ont examiné des chercheurs de Harvard et du MIT : certains groupes se font vacciner, d'autres non. Au premier abord, cette disparité semblait être corrélée avec le budget consacré aux dépenses de santé par les citoyens. En y regardant de plus près, cependant, la corrélation ne se présente pas sous la forme d'une ligne droite, mais plutôt comme celle d'une courbe bizarre. Plus les dépenses de santé voient leur montant s'élever, plus la disparité en matière de vaccination diminue (comme on peut s'y attendre) mais, fait étonnant, si les dépenses augmentent encore plus, cette disparité s'accroît – certains parmi les très riches semblent refuser de se faire vacciner. Pour les responsables en santé publique, c'est un fait primordial à connaître, et une simple analyse de corrélations linéaires n'aurait pu le détecter.

Actuellement, les experts élaborent les outils nécessaires à l'identification et à la comparaison des corrélations non linéaires en même temps qu'on améliore les techniques de l'analyse de corrélations. On fait appel pour cela à des approches innovantes et à de nouveaux logiciels capables d'examiner les relations entre données sous de nombreux angles différents – un peu à la manière des peintres cubistes qui tentaient de capturer l'image d'un visage féminin sous plusieurs angles à la fois. L'une des nouvelles méthodes les plus dynamiques s'applique au domaine en plein essor dit d'analyse des réseaux. Elle permet de faire le recensement, la mesure et divers calculs sur les nœuds de réseau ainsi que sur les liens dans toutes sortes de contextes : avec nos amis sur Facebook, entre les décisions de justice et les citations de jurisprudence, les échanges particuliers sur téléphone portable. Tous ces outils contribuent à fournir les réponses à des questions empiriques et non causales.

À l'ère des big data, ces analyses d'un genre nouveau conduiront au bout du compte à faire surgir des connaissances nouvelles et des prédictions utiles. Des liaisons jusqu'alors invisibles vont apparaître. Il sera possible d'appréhender des dynamiques techniques et sociales complexes qui nous

avaient longtemps échappé malgré tous nos efforts. Surtout, et c'est là l'essentiel, ces analyses non causales nous aideront à mieux comprendre le monde, principalement en posant la question du quoi ? plutôt que du pourquoi ?

Cela peut sembler paradoxal de prime abord. Après tout, en tant qu'humains, nous voulons expliquer le monde en termes de relations de cause à effet ; nous voulons croire qu'il suffit d'y regarder de suffisamment près pour découvrir que tout effet a une cause. Connaître les raisons qui sous-tendent le monde, n'est-ce pas cela qui devrait être notre plus haute aspiration ?

Certes, le débat philosophique sur la causalité – existe-t-elle ou pas ? – dure depuis des siècles. Si tout était processus causal, la logique voudrait que nous n'ayons aucun libre arbitre. La volonté humaine n'existerait pas, du fait que derrière toute décision que nous prendrions et toute pensée qui nous traverserait se cacherait un facteur causal qui, lui aussi, serait la résultante d'une autre cause, et ainsi de suite. Toute vie verrait sa trajectoire déterminée par des causes menant à des effets. D'où les querelles philosophiques sur le rôle de la causalité dans notre monde, et sa possible opposition avec la notion de libre arbitre. Néanmoins, ce débat abstrait n'est pas notre propos.

Quand nous disons que les humains voient le monde à travers le prisme des causalités, nous parlons des deux modes fondamentaux d'explication et de compréhension du monde : la causalité rapide et illusoire et les expériences causales lentes et méthodiques. Les big data vont transformer leur rôle à toutes deux.

Il y a au départ notre désir intuitif de trouver des relations de cause à effet. Nous avons même tendance à en supposer l'existence quand ce n'est pas le cas. Cela ne s'explique ni par la culture, ni par l'éducation ni par le niveau d'instruction. Certaines recherches montrent que c'est le fonctionnement même de la cognition qui nous y conduit. La succession de deux événements pousse notre esprit à les envisager en termes de processus causal.

Prenez les trois phrases suivantes : « Les parents de Fred sont arrivés tard. Les traiteurs étaient attendus sous peu. Fred était furieux. » À la lecture, notre intuition nous dit instantanément que la colère de Fred est due non pas au fait que les traiteurs devaient arriver sous peu, mais au retard de

ses parents. En réalité, nous n'avons pas les moyens d'en être certains à partir de ce qui a été dit. Et pourtant, notre esprit ne peut s'empêcher de créer des histoires que nous tenons pour cohérentes, avec des relations de cause à effet, à partir des faits qui nous sont communiqués.

Daniel Kahneman, professeur de psychologie à Princeton et lauréat en 2002 du prix Nobel d'économie, utilise cet exemple pour suggérer que nous avons deux modes de pensée. L'un, rapide et demandant peu d'effort, nous fait tirer des conclusions hâtives. L'autre, lent et laborieux, exige de nous une concentration particulière.

On a trop tendance, quand on réfléchit vite, à « voir » des liens de causalité même là où il n'y en a pas et à vouloir confirmer nos connaissances et croyances. Autrefois, cette façon rapide de réfléchir nous a bien servis dans un environnement dangereux, quand il nous fallait vite prendre des décisions tout en disposant de peu d'informations. Mais bien souvent, c'est insuffisant pour que soit attribuée à un effet sa véritable cause.

Malheureusement, comme l'explique Kahneman, nous avons un cerveau trop paresseux pour s'adonner à une réflexion lente et méthodique. Le mode de réflexion rapide, auquel nous laissons libre cours, nous conduit à « voir » des processus causaux imaginaires, d'où notre incompréhension fondamentale du monde.

Les parents disent souvent à leurs rejetons qu'ils ont attrapé la grippe parce qu'ils étaient sortis dans le froid sans chapeau ni gants, alors qu'il n'y a pas de lien causal direct entre le fait de s'emmitoufler et d'attraper la grippe. Si on tombe malade après être allé au restaurant, intuitivement on met cela sur le compte de ce qu'on y a mangé (et nous éviterons probablement ce restaurant à l'avenir), même si notre maladie n'a aucun rapport avec la nourriture avalée. Notre problème intestinal a pu avoir de multiples origines, par exemple on a serré la main à une personne infectée. La réflexion rapide de notre cerveau est programmée de manière à sauter rapidement sur une première conclusion de type causal... ce qui, souvent, nous conduit à prendre de mauvaises décisions.

Contrairement aux idées reçues, cette notion de causalité typique de l'humain et qui naît de façon intuitive ne nous aide pas à approfondir notre compréhension du monde. Il ne s'agit souvent que d'un raccourci cognitif qui donne l'illusion de comprendre mais nous maintient en fait dans l'ignorance du monde environnant. À l'instar de l'échantillonnage,

raccourci utile quand on ne peut pas traiter toutes les données, la causalité est un raccourci de la perception, qui permet à notre cerveau de s'éviter une réflexion plus approfondie et chronophage.

Dans un monde de petites données, il a fallu du temps pour démontrer que les intuitions causales étaient une erreur. Cela va changer. Les big data vont régulièrement venir les infirmer en montrant, grâce aux corrélations, qu'il y a peu ou pas de rapport en termes statistiques entre l'effet et sa supposée cause. Résultat : notre mode de « pensée rapide » va profondément se confronter avec la réalité, et pour longtemps.

Nous allons certainement être poussés à réfléchir de façon plus approfondie (et à un rythme plus lent) si nous voulons comprendre le monde. Mais même cette lente réflexion que nous pouvons avoir pour vérifier les processus causaux verra son rôle transformé par les corrélations issues des big data.

Dans notre vie de tous les jours, nous pensons si souvent en termes de causalité que nous en arrivons même à croire qu'on peut aisément en faire la démonstration. En réalité, ce n'est pas facile. À l'inverse des corrélations, et de leurs mathématiques relativement simples, la preuve mathématique de la causalité n'est pas évidente à apporter. Les relations de cause à effet ne s'expriment pas simplement sous la forme d'équations classiques. Même avec une réflexion approfondie, il n'est pas évident d'arriver à la conclusion qu'on les a véritablement mises au jour. Habités que nous sommes à raisonner dans un environnement pauvre en informations, nous nous contentons de données limitées alors que les facteurs en jeu sont bien plus nombreux et qu'on ne saurait réduire un effet à une simple cause.

Prenez le cas du vaccin contre la rage. Le 6 juillet 1885, un garçon âgé de neuf ans, Joseph Meister, qui avait été attaqué par un chien enragé est présenté à Louis Pasteur. Le chimiste français avait inventé la vaccination et travaillé à l'élaboration d'un vaccin expérimental contre la rage. Les parents du garçon supplient Pasteur de recourir au vaccin pour soigner leur fils, ce qu'il fait. Joseph Meister a survécu. Dans la presse, Pasteur est célébré comme le sauveur du jeune garçon, lui ayant épargné une mort certaine et douloureuse.

Mais est-ce bien la vérité ? Comme les faits l'ont montré, sur sept personnes mordues par un chien enragé, une en moyenne contracte effectivement la maladie. À supposer même que le vaccin expérimental de

Pasteur ait été efficace, le garçon avait de toute façon 85 % de chances de survie.

Dans cet exemple, on a considéré qu'il fallait voir dans l'inoculation du vaccin la cause de la guérison de Joseph Meister. Mais en réalité, il faut prendre en compte deux liens de causalité : l'un entre le vaccin et le virus de la rage, et l'autre entre la morsure par un chien enragé et le développement de la maladie. Même si le premier est vrai, le second ne l'est que dans une minorité de cas.

Pour surmonter les difficultés de démonstration d'un processus causal, les scientifiques procèdent à des expériences, dans lesquelles la cause supposée peut être soit introduite soit éliminée. En fonction des effets correspondant respectivement à sa présence ou à son absence, le lien de causalité est établi ou non. Meilleur est le contrôle des conditions expérimentales, plus il est probable que la découverte d'un lien de causalité soit avérée.

Mais il est rare que la preuve de la causalité puisse être apportée de manière irréfutable : elle ne peut l'être qu'à un certain niveau élevé de probabilité. À la différence des corrélations, les expériences à mener ne sont pas pratiques ou soulèvent de sérieux problèmes éthiques. Prenons le cas de la grippe. Quelle expérience pour apporter la preuve que les meilleurs indicateurs de son arrivée correspondent à certaines requêtes et pas à d'autres ? Quant à la vaccination contre la rage, voudrions-nous apporter la preuve de son efficacité en soumettant à une mort douloureuse des dizaines, voire des centaines de patients – ceux du « groupe témoin » qui ne recevrait pas d'injection – alors que nous avons du vaccin à leur disposition ?

Même pour des expériences pratiques, le processus demeure long et coûteux. En comparaison, les analyses non de causalité mais de corrélation sont souvent rapides et économiques. Contrairement à ce qui se passe pour les liens de causalité, les méthodes mathématiques et statistiques existent qui permettent d'analyser les relations entre les données. De même que les outils numériques pour apporter la preuve de la solidité de ces corrélations.

Ces dernières ne sont pas seulement utiles en soi, elles montrent aussi la voie pour une recherche ultérieure des liens de causalité. Si un lien potentiel est trouvé entre deux éléments, cela permet d'approfondir la recherche d'une relation de cause à effet, existante ou non. Et si oui, pourquoi. Des expériences rigoureusement contrôlées permettent ainsi de procéder à un filtrage peu coûteux et très rapide qui diminue le coût de l'analyse causale.

Par le biais des corrélations, il est possible d'obtenir un aperçu des variables à utiliser ensuite pour mener des expériences de causalité.

Cependant, il faut rester prudent. Le rôle que jouent les corrélations est important non seulement parce qu'elles fournissent des éléments d'information, mais aussi parce que le sens en est relativement clair. Chose qui disparaît souvent quand intervient à nouveau la notion de causalité. Exemple, ce concours portant sur la qualité des voitures d'occasion, lancé en 2012 par Kaggle, une société qui organise des concours d'extraction de données, ouverts à tous. L'idée était de créer un algorithme capable de prédire les problèmes que risquaient de présenter les voitures lors d'une vente aux enchères, les statisticiens participant au concours se fondant sur les données fournies par un vendeur de voitures d'occasion. Résultat d'une analyse par corrélation : les voitures de carrosserie orange avaient beaucoup moins tendance à présenter des défauts – près de moitié moins que la moyenne.

Aussitôt, nous pouvons tenter d'en imaginer la raison : se pourrait-il que les propriétaires de voitures de couleur orange soient des passionnés qui entretiennent mieux leur véhicule ? Ou bien que la voiture a été fabriquée avec un soin particulier puisqu'il s'agit d'une couleur personnalisée – la voiture pouvant d'ailleurs être personnalisée sur d'autres plans ? Ou alors, que les voitures de couleur orange se voient mieux sur la route, ont donc moins de risque d'accident et, au moment de la revente, sont en meilleur état ?

Nous nous retrouvons vite piégés dans un filet d'hypothèses contradictoires. En tentant de faire la lumière sur le sujet, on ne fait que le rendre plus confus. Les corrélations existent ; nous en avons la démonstration mathématique, mais il est impossible d'établir de la même manière les liens de causalité. Nous avons donc tout intérêt à nous abstenir de chercher la raison qui se cache derrière les corrélations : le pourquoi au lieu du quoi. Sinon, nous en viendrions à conseiller de peindre sa voiture en orange pour minimiser les défauts du moteur – idée ridicule s'il en est !

Le court exposé précédent montre la supériorité de l'analyse de données fiables par corrélations (ou avec des méthodes non causales du même genre), sur la recherche de relations causales, issues d'une « réflexion rapide » et intuitive. Il se trouve que, dans de plus en plus de cas, cette analyse vaut mieux qu'une longue réflexion en termes de causalité, qui nécessite de longues et coûteuses expériences, soigneusement contrôlées.

Pour en réduire les coûts, ces derniers temps, les scientifiques ont imaginé de mettre sur pied des « quasi-expériences », en associant intelligemment des enquêtes adéquates à leurs données. Une démarche qui peut faciliter la découverte de liens de cause à effet. Mais, en termes d'efficacité, les méthodes non causales sont imbattables ! Les big data commencent par guider les experts vers les causes probables. Et ensuite, ils peuvent approfondir leur travail et rechercher précisément le pourquoi des choses.

La notion de causalité n'est pas abandonnée mais elle perd toutefois son statut, elle qui était considérée comme la source première de sens. Les enquêtes en termes de causalité laissent désormais la place aux big data et aux analyses non causales. En voici une bonne illustration : celle des explosions dans les trous d'homme^e à Manhattan, une énigme jusqu'alors inexpliquée.

Homme contre trou d'homme

Tous les ans, à New York, des centaines de panaches de fumée s'échappent de trous d'homme où des incendies se déclarent. Il arrive que les plaques en fonte de plus de 100 kilos qui couvrent ces trous explosent. Des blocs sont alors éjectés sur une hauteur de plusieurs étages avant de retomber sur le sol, ce qui peut avoir une incidence fâcheuse !

Tous les ans, Con Edison, l'entreprise de services publics qui fournit la ville en électricité, procède à une inspection de ces trous d'homme et y effectue la maintenance nécessaire. Autrefois, elle comptait essentiellement sur le hasard, en espérant que la visite de contrôle programmée d'un trou d'homme aurait lieu avant toute explosion. En 2007, dans le but de concentrer ses effectifs à meilleur escient sur les trous d'homme qui pouvaient présenter un danger, Con Edison s'adressa à des statisticiens de l'université Columbia. Dans ses archives, sont en effet conservées les données historiques concernant le réseau, ses problèmes au cours des années et les connexions entre infrastructures.

Il s'agissait d'un problème complexe de big data. La ville de New York compte 151 278 kilomètres de câbles souterrains, soit trois fois et demie le tour de la Terre. À elle seule, la presque île de Manhattan compte près de 51 000 trous d'homme et coffrets de branchement. Une partie de cette

infrastructure remonte à l'époque de Thomas Edison, qui a donné son nom à la compagnie. Un câble sur vingt a été posé avant 1930. Des archives qui sont conservées depuis les années 1880 se présentaient sous toutes sortes de formes disparates – et il n'avait jamais été envisagé d'y faire appel dans le cadre d'une analyse de données. Elles avaient plusieurs provenances : service comptabilité ; centres d'appels d'urgence avec notes sur les incidents, rédigées à la main par les opérateurs. Dire que les données étaient déstructurées est un euphémisme. À titre d'exemple, les statisticiens ont signalé qu'ils avaient trouvé pour le terme « coffret de branchement^f » (un élément courant de l'infrastructure) au moins 38 appellations différentes, notamment SB, S, S/B, S.B, S?B, S.B., SBX, S/BX, SB/X, S/XB, /SBX, S.BX, S &BX, S?BX, S BX, S/B/X, S BOX, SVBX, SERV BX, SERV-BOX, SERV/BOX et SERVICEBOX. Un algorithme informatique était prié de se débrouiller avec tout ce bazar.

« Les données étaient si incroyablement brutes ! se souvient la statisticienne Cynthia Rudin, spécialiste de l'exploration de données actuellement au MIT, qui a dirigé le projet. Je possède une sortie imprimée de tous les différents tableaux de câbles. Si vous déroulez ce document, il traîne sur le sol même si vous le tenez à bras levé. Et il nous a fallu tout déchiffrer – pour en extraire l'or et mettre tout en œuvre pour qu'un modèle prédictif de très bonne qualité en sorte. »

Pour cette entreprise, Rudin et son équipe ont dû utiliser toutes les données disponibles, et non pas se contenter d'un échantillon, sachant que n'importe lequel des dizaines de milliers de trous d'homme peut constituer une bombe à retardement. Il a fallu faire appel à la formule $N = \text{tous}$. La découverte d'un processus causal dans les incendies de trous d'homme aurait été une bonne chose, certes, mais cela aurait pris un siècle, il y aurait eu des erreurs et tous les problèmes n'auraient pas été pris en compte. Trouver des corrélations est apparu comme la meilleure manière d'accomplir la tâche. Rudin s'est moins préoccupée du pourquoi que du quoi. Ce qu'elle savait, c'est que, face aux dirigeants de Con Edison, l'équipe de cracks en statistique devrait être capable de justifier sur quoi son classement était fondé. Des machines peuvent bien aider à faire les prédictions, mais les commanditaires sont des humains, et les gens en général, pour comprendre un phénomène, veulent en connaître les raisons.

En explorant les données, l'équipe finit par mettre au jour la mine d'or que Rudin espérait trouver. Une fois formatées les données initialement déstructurées, dans le but d'en faire un traitement automatique, l'équipe a

commencé par choisir 106 variables qui peuvent être annonciatrices d'une catastrophe majeure dans un trou d'homme. Elle a poursuivi son travail en réduisant cette liste à un petit nombre de signaux, dont les plus significatifs. Dans un test du réseau électrique du Bronx, elle a analysé toutes les données disponibles allant jusqu'au milieu de l'année 2008. Puis a utilisé ces données pour prédire les emplacements susceptibles de poser problème en 2009. Cela a fonctionné à merveille. Dans les dix pour cent des trous d'homme classés en haut de la liste ont figuré 44 % – soit un pourcentage énorme – de ceux où s'est finalement produit un grave incident.

Au final, quels étaient les facteurs les plus importants ? L'ancienneté des câbles et les antécédents éventuels de problèmes de tel ou tel trou d'homme. Cela tombait bien, en fait, car ainsi il serait facile aux « grosses huiles » de Con Edison de comprendre sur quels critères reposait le classement. Âge et problèmes antérieurs ? Cela ne semble-t-il pas assez évident ? Eh bien, oui et non. D'un côté, comme aime à le dire le théoricien des réseaux Duncan Watts : « Tout est évident, une fois que vous connaissez la réponse » (titre de l'un de ses livres). D'un autre côté, il est important de se rappeler qu'au départ, le modèle comptait 106 variables prédictives. Il n'était pas vraiment évident d'évaluer et classer des dizaines de milliers de trous d'homme, avec pour chacun d'eux d'innombrables variables venant s'ajouter à des millions de points de données. Sans compter qu'au début, les données elles-mêmes ne se présentaient même pas dans un format analysable.

Cet exemple des trous d'homme et de leur risque d'explosion est là pour souligner le fait que l'utilisation des données sert à des fins nouvelles et permet de résoudre des problèmes complexes dans le monde réel. Mais pour y parvenir, il nous a fallu changer de mode opératoire : utiliser toutes les données, toutes celles que nous étions en mesure de collecter, et pas seulement une partie, accepter le désordre et ne plus ériger l'exactitude en première priorité. Enfin, faire confiance aux corrélations sans véritablement connaître la cause derrière les prédictions.

Fin de la théorie ?

Il est clair que les big data transforment notre compréhension du monde et notre façon de l'explorer. Du temps des petites données, on partait d'hypothèses sur le fonctionnement du monde, puis on tentait de les valider

en procédant à la collecte et à l'analyse des données. À l'avenir, c'est l'abondance des données qui va nous donner notre compréhension du monde plus que les hypothèses.

Ces dernières ont souvent été élaborées à partir de théories en sciences naturelles et en sciences sociales, lesquelles contribuent à expliquer et/ou à prédire le monde qui nous entoure. Dans la phase actuelle de transition d'un monde fondé sur la vérification d'hypothèses à un monde axé sur les données, nous pourrions être tentés de penser que nous allons également nous dispenser de théories.

En 2008, le rédacteur en chef du magazine *Wired*, Chris Anderson, a clamé haut et fort : « Le déluge de données rend la méthode scientifique obsolète. » Dans l'article principal intitulé « L'ère du pétaoctet », il proclamait qu'il ne s'agissait de rien de moins que de « la fin de la théorie ». La procédure classique de la découverte scientifique – celle d'une hypothèse qui se confronte à la réalité en faisant appel à un modèle et des causalités sous-jacentes – est appelée à disparaître, affirmait Anderson. Elle allait être remplacée par l'analyse statistique de pures corrélations, d'où la théorie est absente.

À l'appui de cet argument, Anderson décrivait comment la physique quantique était devenue un domaine presque purement théorique, parce que les expériences étaient trop coûteuses, trop complexes et de trop grande envergure pour être réalisables. Il y a la théorie, avançait-il, qui n'a plus rien à voir avec la réalité. Comme exemples de la nouvelle méthode, il citait le moteur de recherche de Google et le séquençage du génome. « C'est un monde où d'énormes masses de données et les mathématiques appliquées remplacent tout autre outil utilisable, écrivait-il. Avec des données suffisantes, les nombres parlent d'eux-mêmes. Avec les pétaoctets, nous pouvons dire : “La corrélation suffit.” »

L'article a provoqué un débat enflammé, même si Anderson est vite revenu sur ses affirmations catégoriques. Mais son argumentaire mérite examen. En substance, Anderson soutient que jusqu'à récemment, pour analyser et comprendre le monde, nous avons besoin de théories à tester. En revanche, à l'ère des big data, toujours selon son argumentation, nous n'avons plus besoin de théories : il nous suffit de regarder les données. Si c'est le cas, alors toutes les règles généralisables relatives au fonctionnement du monde, au comportement humain, aux types d'achats des consommateurs, au moment où une pièce tombe en panne, etc.,

pourraient bien devenir caduques avec l'entrée en piste des big data et de leur analyse.

Parler de « fin de la théorie » semble impliquer que si des domaines de fond comme la physique et la chimie ont eu besoin de théories il n'en irait pas de même avec les big data, dont l'analyse, elle, n'aurait aucun besoin de modèles conceptuels. C'est absurde !

Les big data elles-mêmes ont une base théorique. Par exemple, elles emploient des théories statistiques et mathématiques, et elles utilisent également, à l'occasion, la théorie de l'informatique. Certes, ce ne sont pas des théories sur la dynamique causale d'un phénomène particulier comme la gravité, mais ce sont tout de même des théories. Et, comme nous l'avons montré, les modèles qui les ont pour base possèdent un pouvoir prédictif très utile. En fait, c'est précisément parce qu'elles ne s'encombrent pas de raisonnement classique ni de biais inhérent à telle ou telle théorie que les big data peuvent offrir un regard nouveau et fournir de nouvelles connaissances.

Du fait que l'analyse des big data se fonde sur certaines théories, nous ne pouvons pas échapper à ces dernières. Elles déterminent à la fois nos méthodes et nos résultats. Cela commence par notre façon de choisir les données. Nos décisions peuvent être motivées par un aspect pratique : les données sont-elles facilement disponibles ? ou par un aspect économique : les données peuvent-elles être saisies à faible coût ? Les théories influencent nos choix qui, à leur tour, influent sur les résultats obtenus, comme l'ont soutenu Danah Boyd et Kate Crawford, chercheurs dans le domaine des technologies numériques. Après tout, Google a utilisé des requêtes comme donnée substitutive pour la grippe, et non la longueur des cheveux. De même, quand nous faisons l'analyse des données, nous choisissons des outils dont la conception repose sur certaines théories. Et les théories jouent aussi sur l'interprétation des résultats. C'est évident, les théories ne sont pas absentes de l'ère des big data – mais demeurent omniprésentes, avec tout ce que cela implique.

Anderson a le mérite de soulever les bonnes questions – et de le faire, généralement, avant les autres. Avec les big data, on pourrait bien ne pas arriver à la « fin de la théorie », mais en tout état de cause, elles transforment fondamentalement notre manière d'appréhender le monde. Il faudra du temps pour se faire à ce changement. Elles constituent un défi pour de nombreuses institutions. Mais en raison de la valeur considérable

qu'elles dégagent, elles vont constituer un compromis non seulement valable, mais aussi inévitable.

Cependant, avant d'atteindre ce stade, il est intéressant de se rappeler comment nous en sommes arrivés là. De nombreux professionnels du secteur des technologies se plaisent à mettre cette mutation sur le compte des nouveaux outils numériques – des puces rapides aux logiciels performants – parce que ce sont eux qui les fabriquent. Les prouesses technologiques ont leur importance, mais pas autant qu'on veut bien le dire. La raison profonde de ces évolutions est que nous disposons de bien plus de données. Et cette abondance nouvelle s'explique à son tour par le fait que nous sommes en mesure de convertir au format de données de nombreux aspects de la réalité. Cela sera le sujet du chapitre suivant.

[a.](#) *Le Hasard sauvage. Comment la chance nous trompe*, Les Belles Lettres, Paris, 2009.

[b.](#) Autorité fiscale américaine.

[c.](#) Traduction du terme anglais *Target*.

[d.](#) *Le Pouvoir des habitudes. Changer un rien pour tout changer*, Saint-Simon, Paris, 2013.

[e.](#) Trou permettant le passage d'un homme pour l'inspection et la maintenance d'ouvrages de travaux publics, égouts, etc.
(N.d.É.)

[f.](#) *Service box* (SB) en anglais.

5

Mise en données

Matthew Fontaine Maury, officier de marine de l'U.S. Navy plein d'avenir, avait reçu sa nouvelle affectation à bord du brick Consort en 1839, lorsque sa diligence dérapa brusquement, se renversa et il fut projeté en l'air. Son atterrissage brutal lui valut une fracture du fémur et un genou luxé. Un médecin local remit en place son articulation mais le fémur mal soudé dut être cassé à nouveau quelques jours plus tard. À trente-trois ans, Maury resta partiellement handicapé des suites de ses blessures et inapte à reprendre la mer. Près de trois ans de convalescence plus tard, il fut placé par la Marine dans un bureau, à la tête d'un service au nom peu inspirant : « Dépôt des cartes et instruments de navigation^a. »

Ce poste se révéla parfait pour lui. Jeune navigateur, Maury avait été déconcerté par la façon dont les navires zigzaguaient sur les océans au lieu d'adopter un cap plus direct. Quand il interrogeait les capitaines à ce sujet, ils répondaient qu'il valait bien mieux suivre une route familière que de prendre le risque d'en emprunter une moins habituelle, susceptible de receler des dangers inconnus. Ils considéraient l'océan comme un domaine imprévisible, où les marins étaient confrontés à l'inattendu au moindre vent, à la moindre vague.

Pourtant, à travers ses voyages, Maury apprit que cela n'était pas entièrement vrai. Il observa partout des phénomènes récurrents. Lors d'une escale prolongée à Valparaiso, au Chili, il vit les vents fonctionner comme un mécanisme d'horlogerie. Un vent de fin d'après-midi cessait brusquement au crépuscule et se transformait en une douce brise, comme si quelqu'un avait fermé un robinet. Lors d'un autre voyage, il traversa les eaux bleues du Gulf Stream qui s'écoulait entre les parois sombres des eaux atlantiques, aussi net et régulier que si c'était le fleuve Mississippi. En fait,

les Portugais naviguaient sur l'Atlantique depuis des siècles en se fiant aux vents d'est et vents d'ouest réguliers appelés alizés ; il est à noter que leur nom en anglais est *trades* (ce qui en vieil anglais signifie « chemin » ou « piste » et ne prendra que plus tard le sens de « commerce »).

Chaque fois que l'enseigne de vaisseau Maury arrivait dans un nouveau port, il partait à la recherche de vieux capitaines riches d'expériences transmises de génération en génération pour améliorer ses connaissances à leur contact. Ils lui parlèrent des marées, des vents et des courants marins dont l'action était régulière – mais dont on ne trouvait nulle trace dans les livres et les cartes fournis par la Marine à ses marins. Bien au contraire, ceux-ci se fiaient à des cartes datant parfois de plusieurs siècles et qui comportaient, pour beaucoup d'entre elles, d'énormes omissions ou des inexactitudes flagrantes. Dans ses nouvelles fonctions de surintendant du Dépôt des cartes et instruments de navigation, il se fixa le but d'y remédier.

Dès sa prise de fonctions, il entreprit l'inventaire des baromètres, compas, sextants et chronomètres faisant partie de la collection du dépôt. Il prit également note des myriades de livres nautiques, cartes et cartes marines qu'abritait le dépôt. Il trouva des caisses à l'odeur de moisi pleines de vieux journaux de bord provenant de tous les voyages des capitaines de la Marine. Pour ses prédécesseurs à ce poste, ce n'était que du rebut. Avec quelques courts poèmes ou des dessins dans les marges, ils faisaient parfois plus penser à un moyen d'évasion loin de la traversée qu'à un enregistrement de la localisation des navires.

Mais au fur et à mesure que Maury dépoussiérait les livres tachés par l'eau de mer avant de les ouvrir pour y jeter un coup d'œil, son enthousiasme augmentait. Là étaient les informations qu'il recherchait : des observations sur le vent, sur l'eau et sur le temps dans des zones particulières à des dates spécifiques. Certains des livres de bord présentaient peu d'intérêt, mais beaucoup d'autres fourmillaient d'informations utiles. Si on les rassemblait toutes, comprit Maury, il serait possible de dresser une carte de navigation de conception entièrement nouvelle. Maury et sa douzaine de « calculateurs » – titre des personnes qui faisaient les calculs de données – entamèrent le laborieux processus d'extraction et de compilation des informations enfouies dans les vieux journaux de bord.

Maury agrégea les données et divisa toute la surface de l'océan Atlantique en cases de cinq degrés de longitude et de latitude. Pour chaque case, il consigna la température, la vitesse et la direction des vents et des

vagues, ainsi que le mois, puisque ces conditions diffèrent selon les saisons. En les associant, les données firent ressortir des modèles et donnèrent l'indication d'itinéraires plus efficaces.

Les indications laissées par des générations de marins avaient occasionnellement envoyé des navires directement vers des zones de calme anticyclonique ou les avaient mis aux prises avec des vents et des courants contraires. Sur un itinéraire courant, de New York à Rio de Janeiro, les marins avaient longtemps eu tendance à lutter contre la nature au lieu de s'appuyer sur elle. Pour atteindre Rio, il était conseillé aux capitaines américains d'éviter de naviguer en ligne droite vers le sud, itinéraire à risque. Aussi leurs navires zigzaguaient-ils en mettant cap au sud-est avant de virer au sud-ouest une fois franchi l'équateur. La distance parcourue correspondait bien souvent à trois traversées complètes de l'océan Atlantique. Tous ces détours se révélaient absurdes. Opter pour un « coup quasi direct » ne posait pas de problème.

Pour améliorer la précision, Maury avait besoin de plus amples informations. Il établit un formulaire type destiné à consigner les données relatives aux navires et imposa à tous les vaisseaux de la Marine américaine de l'utiliser et de le remettre à leur arrivée au port. La marine marchande voulait à tout prix se procurer ses cartes marines ; Maury exigea en échange que les navires remettent eux aussi leurs journaux de bord (version ancienne d'un réseau social viral). « Tout navire qui navigue en haute mer, proclama-t-il, peut désormais être considéré comme un observatoire flottant, un temple de la science. » Afin d'améliorer la précision des cartes marines, il rechercha d'autres points de données (tout comme Google s'est appuyé sur l'algorithme PageRank pour inclure plus de signaux). Il obtint des capitaines l'engagement qu'ils lanceraient à la mer, à intervalles réguliers, des bouteilles contenant des notes indiquant le jour, la position, le vent et le courant principal, et qu'ils repêcheraient toute bouteille de ce type qu'ils auraient repérée. De nombreux navires arboraient un pavillon spécial pour indiquer qu'ils coopéraient à cet échange d'informations (préfigurant les icônes de partage de liens qui figurent sur certaines pages Web).

Ces données laissèrent apparaître des voies maritimes naturelles où les vents et les courants étaient particulièrement favorables. Avec les cartes marines de Maury, les longs voyages étaient raccourcis, généralement d'un tiers environ, ce qui faisait économiser beaucoup d'argent aux commerçants. « Avant d'utiliser vos cartes, je traversais l'océan à

l'aveuglette », écrivit un commandant reconnaissant. Et même de vieux loups de mer, qui rejetaient les nouvelles cartes marines pour se fier aux méthodes traditionnelles ou à leur intuition, jouaient un rôle utile : si leurs voyages prenaient plus de temps ou tournaient au désastre, l'utilité du système de Maury était démontrée. En 1855, quand il publia son œuvre magistrale *Géographie physique de la mer*, Maury avait reporté 1,2 million de points de données. « Ainsi, le jeune marin, au lieu d'avancer à tâtons jusqu'à ce que les expériences accumulées éclairent sa voie... y découvrirait, sur-le-champ, qu'il disposait déjà de l'expérience d'un millier de navigateurs pour le guider », écrivit-il.

Son œuvre a joué un rôle essentiel dans la pose du premier câble télégraphique transatlantique. Et, après une tragique collision en haute mer, il conçut en hâte le système de voies maritimes couramment utilisé aujourd'hui. Il appliqua même sa méthode à l'astronomie : à la découverte de la planète Neptune en 1846, Maury eut la géniale idée de ratisser les archives à la recherche de documents où elle était désignée à tort comme une étoile, ce qui permit de tracer son orbite.

Il a été généralement fait peu de cas de Maury dans les livres d'histoire américains, sans doute parce que ce natif de l'État de Virginie démissionna de l'Union Navy^b durant la guerre de Sécession et devint espion en Angleterre pour le compte des confédérés. Mais des années plus tôt, quand il vint en Europe pour obtenir un soutien international à ses cartes marines, quatre pays le firent chevalier et il reçut des médailles d'or dans huit autres, notamment au Saint-Siège. À l'aube du XXI^e siècle, des cartes de pilotage publiées par la Marine américaine portaient encore son nom.

Le commandant Maury, l'« Éclaireur des mers », fut un des premiers à comprendre la valeur spéciale d'un immense corpus de données comparé à de plus petites quantités – principe de base des big data. Plus fondamentalement, il avait compris que de véritables « données » pouvaient être extraites et compilées à partir des journaux de bord moisissés de la Marine. Grâce à cela, il fut l'un des pionniers de la « mise en données », de la mise au jour de données à partir de documents dont personne ne soupçonnait la valeur. À l'instar d'Oren Etzioni chez Farecast, qui a développé une activité lucrative en utilisant des informations portant sur d'anciens tarifs de compagnies aériennes, ou des ingénieurs de Google, qui ont suivi les épidémies de grippe en se servant de vieilles requêtes, Maury

s'est servi d'informations produites dans un but déterminé et les a converties en quelque chose d'autre.

Sa méthode, largement similaire aux techniques actuelles des big data, était stupéfiante sachant que les seuls outils d'alors étaient le crayon et le papier. Avec son histoire, on comprend que l'utilisation des données est antérieure à la numérisation. Aujourd'hui, nous avons tendance à confondre les deux, alors qu'il est important de les séparer. Pour mieux saisir la manière dont des données peuvent être extraites des sources les plus improbables, prenons un exemple plus moderne.

Étudier le postérieur de ses congénères, tels sont l'art et la science pratiqués par Shigeomi Koshimizu, professeur à l'université de l'Institut avancé de technologie industrielle à Tokyo. Rares sont ceux qui imagineraient que la façon dont quelqu'un s'assoit constitue une information, mais c'est le cas. Quand quelqu'un est assis, les contours du corps, la posture et la répartition du poids peuvent tous être quantifiés et compilés. Koshimizu et son équipe d'ingénieurs mesurent avec des capteurs la pression exercée en 360 points sur un siège de voiture et indexent chaque point sur une échelle de zéro à 256 pour obtenir au final un code numérique unique pour chaque personne. Ce qui convertit tous ces postérieurs en données. Lors d'un essai, le système a pu distinguer les personnes d'un petit groupe avec 98 % d'exactitude.

Cette recherche est tout sauf gratuite. Elle est destinée à la mise au point d'un système d'antivol sur les voitures. Un véhicule équipé de ce dispositif pourrait reconnaître la personne au volant et s'il s'apercevait qu'il ne s'agit pas du conducteur agréé, il lui demanderait un mot de passe et, selon le cas, le laisserait conduire ou bloquerait le moteur. Il pourrait donc non seulement prévenir le vol d'un véhicule mais aussi identifier le voleur « par derrière » (si l'on ose dire). La transformation de positions assises en données s'inscrit dans la conception d'un service utile et la naissance d'une activité potentiellement lucrative. Son utilité pourrait aller bien au-delà de la dissuasion contre le vol de voiture. Par exemple, à partir des données agrégées, on pourrait trouver des indices sur la relation entre la posture du conducteur et la sécurité routière, comme un changement de position révélateur avant un accident. Le système pourrait détecter la fatigue du conducteur, la somnolence entraînant un léger affaissement du corps, et émettre alors un signal d'alerte ou freiner automatiquement.

Le professeur Koshimizu a pris comme objet d'étude et converti en format numériquement quantifié un élément qui n'avait jamais été traité comme une donnée. Personne n'ayant même imaginé que cette information pouvait avoir un sens. De même, le commandant Maury avait transformé en données éminemment utiles des informations extraites d'un matériel sans grand intérêt apparent. Ce faisant, il avait réussi à créer une valeur inédite, en utilisant les informations d'une manière nouvelle.

Le terme de « data » signifie « données » en latin, au sens d'un « fait ». C'est le titre d'un ouvrage classique d'Euclide, dans lequel il explique la géométrie à partir de ce qui est connu et attesté comme tel. Aujourd'hui, le terme de « données » renvoie à la description de quelque chose qui se laisse enregistrer, analyser et réorganiser. Il n'existe pas encore de terme adéquat pour les transformations particulières qu'effectuent des gens comme le commandant Maury ou le professeur Koshimizu. Aussi les appellerons-nous « mise en données ». Mettre en données un phénomène consiste à le transposer en un certain nombre de chiffres que l'on va pouvoir mettre en tableau, puis analyser.

Encore une fois, ce processus est distinct de la numérisation, qui consiste à convertir des informations analogiques en code binaire composé de zéros et de uns, que les ordinateurs savent manipuler. La numérisation n'a pas été la première opération accomplie grâce à l'ordinateur. Dans les débuts de la révolution informatique, il servait au calcul, comme le suggère l'étymologie du terme anglais *computer*^e. Nous utilisions des machines pour effectuer des calculs trop lents à réaliser avec les méthodes antérieures : tableaux de trajectoires de missiles, recensements et climat... La numérisation de contenus analogiques n'est venue que plus tard. C'est pourquoi, quand Nicholas Negroponte du MIT Media Lab a publié en 1995 son ouvrage qui a fait date, intitulé *L'Homme numérique*, un de ses grands thèmes était le passage des atomes aux octets. Nous avons amplement numérisé des textes dans les années 1990. Plus récemment, avec l'augmentation de la capacité de stockage, de la puissance de calcul et l'élargissement de la bande passante, nous l'avons aussi fait avec d'autres contenus : images, vidéos et musique notamment.

Aujourd'hui, les technologues sont en général implicitement convaincus que l'origine des big data remonte à la révolution du silicium. Ce n'est absolument pas le cas. Certes, les systèmes informatiques modernes ont rendu possible leur utilisation, mais fondamentalement, le

passage aux larges volumes de données se situe dans le prolongement de la quête ancestrale propre à l'humanité, celle de mesurer, enregistrer et analyser le monde. La révolution des TI est visible tout autour de nous, mais l'accent a été essentiellement mis sur le T, la technologie. Il est temps d'y jeter un nouveau regard en portant notre attention sur le I, l'information.

Pour recueillir des informations quantifiables et les mettre en données, nous devons savoir comment mesurer et comment enregistrer ce dont nous effectuons la mesure. Et pour cela, il faut, outre des outils adaptés, avoir la volonté de quantifier et d'enregistrer – deux conditions préalables à la mise en données. Ces éléments fondamentaux nécessaires à la mise en données, nous les avons élaborés voilà de nombreux siècles, bien avant l'aube de l'ère numérique.

Quantifier le monde

La capacité à enregistrer les informations constitue l'une des lignes de démarcation entre sociétés primitives et évoluées. Le calcul élémentaire et la mesure de la longueur et du poids font partie des plus vieux outils conceptuels des anciennes civilisations. Au troisième millénaire av. J.-C., d'importants progrès avaient été faits dans la vallée de l'Indus, en Égypte et en Mésopotamie dans l'art de consigner les informations. La précision s'est alors améliorée, tout comme l'usage de la mesure dans la vie quotidienne. L'écriture a évolué en Mésopotamie de manière à conserver de façon précise la trace de la production et des transactions commerciales. Le langage écrit a permis à ces civilisations d'effectuer des mesures de la réalité, de les consigner et d'en retrouver la trace ultérieurement. Ces deux actions – mesurer et consigner – ont facilité la création de données et ont posé les fondations de la mise en données.

C'est ce qui a permis aux humains de reproduire leurs créations dans différents domaines. On pouvait faire des répliques de bâtiments, par exemple, grâce aux informations qui avaient été consignées : dimensions, matériaux utilisés. Cela a aussi permis l'expérimentation : un architecte ou un maçon pouvait concevoir un nouveau bâtiment en modifiant certaines dimensions, tout en en gardant d'autres inchangées – ce qui pouvait à son tour être consigné. On pouvait noter les transactions commerciales et connaître le rendement d'une culture à partir du volume d'une récolte de fruits ou de céréales (et la proportion qui serait prélevée par l'État sous

forme d'impôts). La quantification a permis la prévision et donc la planification, même de façon approximative : s'attendre par exemple simplement à ce que la récolte d'une année soit aussi abondante que celle des années précédentes. Les partenaires étaient en mesure de conserver une trace de leur transaction et de ce que l'un devait à l'autre. Sans la capacité à mesurer et à consigner, l'argent n'aurait pu exister, car il n'y aurait pas eu de données pour en conserver la trace.

Au cours des siècles, les hommes ont élargi le champ de la mesure : à la longueur et au poids sont venus s'ajouter la superficie, le volume et le temps. Au début du premier millénaire av. J.-C., les principaux éléments de mesure étaient désormais établis en Occident. Reste que les méthodes de mesure du temps présentaient une importante lacune. Elles n'étaient pas optimisées pour le calcul, pas même pour un calcul relativement simple. Le système de comptage des chiffres romains était mal adapté à l'analyse numérique. Sans un système de numérotation « positionnel » à dix chiffres ou décimales, la multiplication et la division de grands nombres étaient des opérations difficiles à réaliser, même pour des experts, et les plus simples comme l'addition et la soustraction n'étaient pas évidentes non plus pour la plupart des non-spécialistes.

Un système de chiffres différent a été conçu en Inde vers le 1^{er} siècle av. J.-C. avant de passer en Perse, où il a été amélioré puis transmis aux Arabes qui l'ont grandement affiné. Il constitue la base des chiffres arabes que nous utilisons aujourd'hui. Les croisades ont sans doute apporté leur lot de destruction sur les terres envahies par les Européens, mais le savoir, lui, a migré d'est en ouest, et les chiffres arabes ont sans doute constitué son apport majeur. Après les avoir étudiés, le pape Sylvestre II a préconisé leur utilisation à la fin du premier millénaire. Au XII^e siècle, les textes arabes qui décrivaient le système ont été traduits en latin puis diffusés dans toute l'Europe. C'est ainsi que les mathématiques ont pris leur essor.

Même avant l'arrivée des chiffres arabes en Europe, le calcul avait été facilité par l'utilisation des abaques – des tables à calcul en forme de plateaux lisses sur lesquels étaient placés des jetons indiquant certains montants. On pouvait faire des additions ou des soustractions en faisant glisser les jetons dans certaines zones du plateau, mais la méthode était très limitée. Il était difficile de faire en même temps des calculs sur de très grands nombres et sur de très petits. Point très important, les nombres placés sur les tables n'étaient pas fixés. Un chiffre pouvait se trouver

modifié lors d'un faux mouvement ou d'une secousse imprudente et conduire à des résultats incorrects. Si les abaquages facilitaient le calcul, elles ne convenaient pas du tout pour consigner les nombres et le seul moyen d'y parvenir, puis de les stocker, était de les traduire en chiffres romains, méthode absolument inefficace. (Les Européens n'eurent jamais l'usage des bouliers d'Orient et, avec le recul, c'est une bonne chose, car avec eux, l'usage des chiffres romains se serait sans doute perpétué en Occident.)

Les mathématiques ont conféré un nouveau sens aux données – on pouvait maintenant en faire l'analyse, et pas seulement les consigner et les récupérer. Il a fallu des centaines d'années, depuis leur introduction au XII^e siècle jusqu'à la fin du XVI^e siècle pour que les chiffres arabes soient adoptés en Europe de façon généralisée, et que les mathématiciens se targuent d'être capables de calculer six fois plus vite avec les chiffres arabes qu'avec les abaquages. Mais c'est finalement un autre outil de mise en données dont l'évolution a contribué au succès des chiffres arabes : la comptabilité en partie double.

Les comptables ont inventé l'écriture au troisième millénaire av. J.-C. Même si elle a changé au cours des siècles suivants, la comptabilité est essentiellement demeurée un système d'enregistrement des transactions. En revanche, ce qu'il lui était impossible de faire apparaître, avec facilité et au moment voulu, c'était l'information la plus importante aux yeux des comptables et de leurs employeurs : la rentabilité d'un compte particulier ou de toute une entreprise. Les choses ont commencé à changer au XIV^e siècle quand, en Italie, les comptables se mirent à utiliser deux entrées pour enregistrer les transactions, l'une pour les crédits, l'autre pour les débits, et ainsi équilibrer la balance d'un compte. La merveille avec ce système, c'est que les bénéfices et les pertes apparaissaient clairement. Et soudain, les données si ternes jusque-là se mirent à parler.

De nos jours, ce sont la comptabilité et la finance qui continuent de prendre en compte la tenue de livres en partie double. Mais cette méthode a représenté un jalon décisif dans l'évolution de l'utilisation des données. Elle a permis de créer des catégories d'informations, reliant les comptes. De consigner les données d'une manière codifiée – un des premiers exemples d'enregistrement d'informations au moyen d'un ensemble de règles normalisées. Un comptable pouvait regarder les livres de comptes d'un confrère et les comprendre. Elle a été organisée de manière à pouvoir effectuer rapidement et simplement un type particulier de requête – calculer

les bénéfices ou les pertes relatifs à chaque compte. Elle a fourni une piste pour faire un audit des transactions et faciliter la traçabilité des données. Les mordus de la technologie peuvent encore l'apprécier de nos jours : elle intégrait une fonction de « correction d'erreurs ». Si quelque chose semblait clocher dans une colonne du grand livre, il était possible de vérifier l'entrée correspondante.

Toutefois, à l'instar des chiffres arabes, la comptabilité en partie double ne rencontra pas un succès immédiat. Il a fallu attendre deux siècles avant qu'un mathématicien et une famille de commerçants ne changent le cours de l'histoire de la mise en données.

Le premier était un moine franciscain du nom de Luca Pacioli. En 1494, il publie un manuel de mathématiques à l'intention des profanes, qui remporte un vif succès et devient *de facto* le manuel d'application commerciale des mathématiques de son époque. C'est également le premier livre à n'utiliser que les chiffres arabes, et leur adoption en Europe se généralise grâce à ce succès. Mais sa contribution la plus durable viendra du chapitre consacré à la tenue des livres, où Pacioli explique avec précision le système de comptabilité à double entrée. Cette partie sera publiée en six langues dans les décennies suivantes, et il demeura l'ouvrage de référence sur le sujet pendant des siècles.

Quant à la famille de commerçants, il s'agit des Médicis, les célèbres marchands et mécènes vénitiens. Au XVI^e siècle, notamment parce qu'ils utilisent cette méthode performante d'enregistrement de données qu'est la tenue des livres en partie double, ils deviennent les banquiers les plus influents d'Europe. Ensemble, Pacioli et les Médicis scellent la victoire de ce type de comptabilité – et instaurent l'usage des chiffres arabes en Occident.

En parallèle, les différentes méthodes de mesure du monde gagnent en précision – mesure du temps, de la distance, de la superficie, du volume et du poids. Au XIX^e siècle, la science et sa soif de compréhension du monde se caractérisent par la quantification : les inventions et unités de mesure se multiplient pour consigner les mesures de courant électrique, la pression atmosphérique, la température, les fréquences sonores, etc. À cette époque, tout doit être absolument défini, délimité et noté. L'obsession va même jusqu'à la mesure des dimensions du crâne humain censées indiquer les facultés cognitives de l'individu. Fort heureusement, la pseudoscience que

fut la phrénologie a pratiquement disparu, mais la volonté de quantifier, elle, n'a fait que croître et embellir.

La mesure de la réalité et l'enregistrement des données ont vu leur développement favorisé par deux facteurs : les outils mis en œuvre et l'ouverture d'esprit. C'est la combinaison des deux qui a créé un terreau fertile à la mise en données moderne. Ses ingrédients étaient en place, même si dans le monde de l'analogique, elle est demeurée chère et chronophage. Dans de nombreux cas, elle a même exigé une patience infinie. Il fallait y consacrer toute une vie – à preuve Tycho Brahe et ses fastidieuses observations des étoiles et des planètes dans les années 1500. Le succès – même limité à quelques cas – de la mise en données au cours de la période de l'analogique, comme les cartes de navigation du commandant Maury, a souvent été dû à une heureuse coïncidence : Maury, par exemple, qui fut relégué derrière un bureau se retrouva avoir accès à ce trésor que représentaient des journaux de bord. Et toute opération réussie de mise en données aura été productrice d'une immense valeur ajoutée, et aura jeté un formidable éclairage sur les informations initiales.

Son efficacité a été démultipliée par l'arrivée des ordinateurs, des appareils de mesure numériques et des dispositifs de stockage. La valeur insoupçonnée de l'analyse mathématique s'est révélée. En bref, la performance de la mise en données s'est accrue avec la numérisation. Mais il ne faut pas y voir un substitut. L'opération en soi de numérisation – c'est-à-dire la transformation d'une information analogique en un format lisible par l'ordinateur – n'entraîne pas automatiquement une mise en données.

Quand les mots deviennent des données

La différence entre les deux devient évidente si on observe un domaine qui a connu les deux opérations, dont nous pouvons comparer les conséquences – les livres par exemple. En 2004, Google a fait l'annonce d'un projet incroyablement audacieux. Le contenu intégral de tous les livres auxquels l'entreprise pourrait avoir accès (dans la mesure où ils seraient libres de droit) serait mis sur son site, ce qui permettrait au monde entier de les consulter gracieusement sur Internet. Pour réaliser une telle prouesse sur des millions de livres, Google s'est associé avec certaines des plus grandes et plus prestigieuses bibliothèques universitaires et a mis au point des

scanners capables de tourner automatiquement les pages de manière que l'opération soit à la fois faisable et viable financièrement.

Google a commencé par numériser le texte : chaque page a été scannée puis enregistrée sous forme de fichier image à haute résolution, et enfin stockée sur un serveur. La page avait été transformée en une copie numérique facile à retrouver n'importe où par n'importe quel navigateur. Problème, pour retrouver telle ou telle information, il aurait fallu savoir quel livre la contenait, et parcourir des textes très volumineux avant de trouver le passage recherché. Il aurait été impossible de rechercher des mots particuliers ni de faire l'analyse du texte parce que ce dernier n'avait pas été mis en données. Tout ce dont disposait Google, c'étaient des images que seul un humain avait la capacité de transformer en information utile – en les lisant.

Cela aurait été, certes, un extraordinaire outil – une version moderne, numérique de la Bibliothèque d'Alexandrie, plus complète que toute autre bibliothèque dans l'histoire – mais l'ambition de Google allait bien au-delà. La société avait compris que seule la mise en données des informations pourrait faire apparaître leur valeur cachée. Pour fournir au final un texte mis en données et non l'image numérisée d'une page, Google a utilisé le logiciel de reconnaissance optique de caractères qui reconnaît les lettres, les mots, les phrases et les paragraphes à partir d'une image numérique.

Les informations figurant sur la page ont alors pu être traitées par ordinateur et analysées à l'aide d'algorithmes, et pas seulement accessibles à un lecteur humain. La mise en données a offert la possibilité d'indexer le texte et d'y faire ainsi des recherches. Elle a permis une analyse textuelle infinie. Il nous est désormais possible de découvrir quand certains mots ou certaines phrases ont été utilisés pour la première fois, ou bien sont devenus populaires. Cela apporte un éclairage nouveau sur la diffusion des idées et l'évolution de la pensée à travers les siècles et dans diverses langues.

Vous pouvez faire le test vous-même. Si vous effectuez une requête sur certains mots ou certaines phrases à l'aide du moteur de recherche Ngram Viewer de Google (<http://books.google.com/ngrams>), vous obtiendrez un graphique précisant l'évolution de leur utilisation à travers les époques ; Ngram Viewer se sert de l'index complet de Google Books comme d'une base de données. En quelques secondes, on découvre que le terme de « causalité » était plus fréquemment utilisé que celui de « corrélation » jusqu'en 1900, puis que l'ordre s'est inversé. On peut comparer les styles

littéraires et comprendre les controverses en paternité de certaines œuvres. On découvre aussi plus facilement le plagiat de certains travaux universitaires, grâce à cette mise en données ; c'est ce qui a contraint, par exemple, un certain nombre d'hommes politiques européens à donner leur démission, dont un ministre allemand de la Défense.

Selon les estimations, 130 millions de titres ont été publiés depuis l'invention de l'imprimerie au milieu du xv^e siècle. En 2012, sept ans après le démarrage de son projet sur les livres, Google a scanné plus de 20 millions de titres, correspondant à plus de 15 % du patrimoine écrit mondial – une part considérable. D'où la naissance d'une nouvelle discipline académique, appelée *culturomics* : il s'agit de lexicologie par informatique s'efforçant de comprendre le comportement humain et les tendances culturelles à travers l'analyse quantitative de textes.

Des chercheurs de Harvard ont ainsi révélé que moins de la moitié des mots anglais apparaissant dans les livres se retrouvent dans le dictionnaire, dans une étude qui a consisté à consulter plus de 500 milliards de mots au travers de plusieurs millions de livres. Il existe ainsi, selon eux, une masse de termes qui forment une sorte de « “matière noire” lexicale, dont on ne retrouve pas de trace dans les références normalisées ». Par ailleurs, les chercheurs ont démontré que la suppression ou la censure d'une idée ou d'une personne laissait des « empreintes quantifiables », après avoir effectué une analyse algorithmique des références à l'artiste Marc Chagall, dont les œuvres ont été interdites en Allemagne au temps du régime nazi parce qu'il était juif. Les mots sont semblables à des fossiles enfouis dans les pages au lieu de l'être dans de la roche sédimentaire. Les adeptes de la « culturomique » peuvent mener leurs explorations à la manière des archéologues. Naturellement, l'ensemble de données comporte d'innombrables biais implicites – les livres qu'on trouve dans les bibliothèques sont-ils le reflet fidèle du monde réel ou simplement celui de ce qui est cher aux auteurs et aux librairies ? Néanmoins, la culturomique nous fournit un prisme entièrement nouveau pour mieux nous comprendre nous-mêmes.

Transformer les mots en données ouvre la voie à de nombreuses utilisations potentielles. Oui, les humains peuvent lire et les machines analyser les données. Mais Google, exemple même de grande société des big data, sait que la collecte et la mise en données des informations ont bien d'autres fins potentielles. Aussi la société s'est-elle attelée à améliorer son

service de traduction automatique en se servant habilement des textes mis en données dans le cadre de son projet de scannage de livres. Comme expliqué dans le chapitre 3, le système a analysé le choix des phrases et des mots que les traducteurs ont fait pour traduire un livre d'une langue dans une autre. Sur cette base, il a alors pu traiter la traduction comme un problème de mathématiques géant, et l'ordinateur calculer les probabilités pour qu'un terme se substitue mieux qu'un autre au mot employé dans l'autre langue.

Bien sûr, Google n'a pas été la seule organisation à rêver d'intégrer dans l'ère informatique le riche patrimoine mondial, et elle n'a pas été la première à s'y lancer. Il y a eu le Projet Gutenberg, une initiative de bénévoles pour mettre en ligne dès 1971 des ouvrages passés dans le domaine public. Mais il ne s'agissait que de mettre des textes à lire à disposition de tous, sans envisager d'utilisations annexes consistant à traiter les mots comme des données. Il s'agissait de lire, et non de réutiliser. De même, cela fait des années que les éditeurs se sont lancés dans la publication de versions électroniques de leurs livres. Eux aussi ont entrevu la valeur fondamentale des livres en tant que contenu, et non en tant que données – leur modèle d'entreprise est fondé là-dessus. D'ailleurs, ils n'ont jamais utilisé ni permis à d'autres d'utiliser les données inhérentes au texte d'un livre. Ils n'en ont jamais vu l'utilité, ni bien évalué le potentiel.

De nombreuses sociétés se font actuellement concurrence pour s'implanter sur le marché du livre électronique. Amazon, avec ses lecteurs de livres numériques Kindle, semble avoir pris une bonne longueur d'avance. Mais c'est un domaine où les stratégies d'Amazon et de Google divergent.

Amazon, elle aussi, met en données des livres – mais à l'inverse de Google, la société n'exploite pas les nouvelles utilisations potentielles des textes sous forme de données. Jeff Bezos, le fondateur et directeur général de la société, a convaincu des centaines d'éditeurs de publier leurs livres en format Kindle. Les livres Kindle ne sont pas constitués d'images de pages. Si c'était le cas, il ne serait pas possible de modifier la taille de la police de caractères ni d'afficher une page sur un écran couleur aussi bien que noir et blanc. Le texte n'est pas seulement numérisé, il est mis en données. Amazon a effectivement fait pour des millions de nouveaux livres ce que Google essaie laborieusement de réaliser pour les ouvrages plus anciens.

Cependant, en dehors du service très performant qu'elle propose de « termes statistiquement importants » (des algorithmes trouvent les liens entre les thèmes de livres qui n'apparaîtraient pas autrement) Amazon ne s'est pas servie de la vaste quantité de mots dont elle dispose pour effectuer une analyse de gros volumes de données. Ce commerçant en ligne considère que son activité commerciale de libraire est basée sur le contenu destiné à être lu, plutôt que sur l'analyse de texte mis en données. Et en toute justice, il se heurte probablement au conservatisme des éditeurs quant à la manière d'utiliser les informations contenues dans leurs livres. Google, du fait de son statut d'enfant terrible des big data qui cherche à repousser les limites, ne sent pas les mêmes contraintes : ses recettes proviennent des clics des utilisateurs, et non de l'accès aux titres des éditeurs. On peut honnêtement dire que, du moins pour le moment, Amazon comprend la valeur de la numérisation du contenu, tandis que Google comprend celle de la mise en données du contenu.

Quand l'emplacement se transforme en données

L'un des éléments d'information les plus fondamentaux au monde est... le monde lui-même. Mais durant la majeure partie de l'Histoire, l'espace et ses différentes zones n'ont jamais fait l'objet de quantifications ni été utilisés sous forme de données. La géolocalisation que ce soit de la nature, des objets ou des personnes constitue bien évidemment une information. Telle montagne est là ; telle personne se trouve ici. Pour en optimiser l'utilisation, cette information doit cependant être transformée en données. Or, la mise en données d'un lieu exige quelques conditions préalables. Il nous faut une méthode pour mesurer chaque centimètre carré sur la surface de la Terre, un moyen normalisé de noter les mesures, un instrument pour contrôler et enregistrer les données. Quantification, normalisation, collecte. C'est seulement à ces conditions qu'il est possible de stocker et analyser un lieu non en soi, mais en tant que données.

En Occident, une telle quantification a débuté avec les Grecs. Vers 200 av. J.-C., Ératosthène inventa un système de grille pour préciser un emplacement, un peu à la manière de la latitude et de la longitude. Mais comme ce fut le cas pour bon nombre de bonnes idées nées dans l'Antiquité, la pratique disparut avec le temps. Un millénaire et demi plus tard, vers 1400 av. J.-C., un exemplaire de l'œuvre de Ptolémée,

Géographie, arriva à Florence en provenance de Constantinople. C'était juste au moment où l'intérêt pour la science et le savoir-faire des Anciens se réveillait, avec la Renaissance et l'essor du commerce maritime. Le traité de Ptolémée fit sensation, et ses anciennes leçons furent appliquées pour résoudre des problèmes de navigation contemporains. Dès lors, les cartes mentionnèrent la longitude, la latitude et l'échelle. Le système bénéficia plus tard d'améliorations apportées par le cartographe flamand Gerardus Mercator en 1570, ce qui permit aux marins de tracer un cap dans un monde sphérique.

À cette époque, il n'existait aucun format accepté par tous permettant d'échanger les informations sur un lieu. Il fallait mettre sur pied un système d'identification commun, tout comme Internet a bénéficié du système de noms de domaines permettant au courriel de fonctionner dans le monde entier. La normalisation de la longitude et de la latitude s'est fait attendre. Jusqu'en 1884 où elle fut définie lors de la Conférence internationale de Washington par 25 nations qui choisirent Greenwich, en Angleterre, comme méridien origine et point de longitude zéro (les Français, qui se considéraient leaders en matière de normes internationales, s'abstinrent lors du vote). Dans les années 1940, fut ensuite créée la projection Transversale universelle de Mercator (UTM), un système de coordonnées qui permet d'obtenir une exactitude plus grande encore en divisant le monde en 60 zones.

Il devenait alors possible d'identifier, d'enregistrer, de calculer, d'analyser et de communiquer en format numérique normalisé un emplacement « géospatial ». La position pouvait être mise en données. Mais dans le contexte de l'analogique, cela a été rarement mis en pratique, en raison du coût élevé de la mesure et de l'enregistrement des informations. Jusqu'aux années 1970, il fallait utiliser des bornes, des constellations astronomiques, la technologie limitée du positionnement par radio. Il fallait donc inventer de nouveaux outils permettant de faire cette mesure à moindre coût.

Un grand changement s'est produit en 1978, quand le premier des 24 satellites qui constituent le Système de positionnement global (GPS) a été lancé. En notant le délai de réception du signal satellite (situé à près de 20 300 kilomètres d'altitude), les récepteurs au sol ont pu déterminer leur position par triangulation. Élaboré par le département de la Défense des États-Unis, le système a été ouvert pour la première fois à des fins non

militaires dans les années 1980 et il est devenu totalement opérationnel dans les années 1990. Sa précision a été améliorée pour des applications commerciales une décennie plus tard. Précis au mètre près, le GPS a marqué le moment où la mesure d'un lieu – rêve des navigateurs, des cartographes et des mathématiciens depuis l'Antiquité – peut enfin être réalisée par une méthode rapide et (relativement) à moindre coût, sans requérir de connaissances spécialisées, grâce à des techniques *ad hoc*.

Encore faut-il pouvoir produire concrètement les informations. Rien n'empêchait Ératosthène ni Mercator d'évaluer leur localisation à toute minute de la journée, s'ils l'avaient voulu. C'était certes réalisable, mais peu pratique. De la même manière, si les premiers récepteurs GPS, complexes et coûteux, convenaient pour un sous-marin, ce n'était pas vrai pour n'importe qui, n'importe quand. L'omniprésence des puces bon marché intégrées à toutes sortes de gadgets numériques a bouleversé la donne. Un module GPS a vu son prix dégringoler de plusieurs centaines de dollars dans les années 1990 à près d'un dollar aujourd'hui. Cela ne lui prend généralement que quelques secondes pour déterminer un emplacement, et les coordonnées sont normalisées. Ainsi, 37° 14' 06 N, 115° 48' 40 W a une signification et une seule, à savoir que l'on se trouve dans une base militaire américaine ultrasecrète au fin fond du Nevada connue sous le nom de « Zone 51 », où des extraterrestres sont (peut-être !) détenus.

Aujourd'hui, le GPS n'est qu'un système parmi de nombreux autres pour identifier un emplacement. Des systèmes satellites concurrents sont en cours d'élaboration en Europe et en Chine. Il est même possible d'obtenir une précision encore plus grande par triangulation du signal dont la force dépend de la position des antennes-relais de téléphone cellulaire ou les routeurs wifi (le GPS ne fonctionne en effet pas à l'intérieur ou au milieu de bâtiments élevés). Cela explique pourquoi des entreprises comme Google, Apple et Microsoft ont créé leur propre système de géolocalisation pour compléter le GPS. Les voitures de Google Street View prennent des photos pendant qu'elles collectent les informations via des routeurs wifi, et l'iPhone est un « SpyPhone^d » qui recueille des informations sur des données et des emplacements wifi qui les renvoie à Apple, à l'insu des utilisateurs. (Les Androidphones de Google et le système d'exploitation mobile de Microsoft collectent eux aussi ce type de données.)

Aujourd'hui, on peut suivre non seulement des personnes, mais aussi des objets. Avec des modules sans fil placés à l'intérieur des véhicules, la mise en données d'un emplacement va transformer le concept de l'assurance. Les risques et le coût afférent d'un trajet automobile seront déterminés plus précisément grâce à l'aperçu « granulaire » qu'en donneront les données (date du trajet, zone traversée, distance parcourue...). Aux États-Unis et en Grande-Bretagne, les automobilistes peuvent désormais bénéficier d'une assurance automobile dont le prix est fixé en fonction du lieu et du moment où ils conduisent effectivement, au lieu de payer un simple taux annuel basé sur leur âge, leur sexe et l'historique de leur conduite. Cette tarification de l'assurance est une méthode qui encourage le bon comportement au volant. Elle transforme la nature même de l'assurance qui se base sur l'action individuelle, au lieu de la mise en commun des risques. Le suivi des gens à partir de leur véhicule modifie également la nature des coûts fixes – routes et autres types d'infrastructures – dont l'utilisation est liée au comportement des automobilistes et autres « consommateurs ».

Chose impossible il y a peu car on ne pouvait pas convertir la géolocalisation en données à tout instant pour tout et tous – mais c'est le monde vers lequel nous nous dirigeons maintenant.

UPS, par exemple, utilise les données de géolocalisation à de multiples fins. Ses véhicules sont équipés de capteurs, de modules sans fil et de GPS de façon que tout problème mécanique puisse être anticipé depuis le siège de la société, comme nous l'avons vu dans le précédent chapitre. Cela lui permet aussi de savoir où se trouvent les camions en cas de retard, et d'optimiser les itinéraires en faisant le suivi de ses employés. Des livraisons antérieures fournissent des données qui permettent en partie de déterminer le trajet le plus efficace, un peu comme les journaux de bord des voyages en mer avaient permis à Maury de tracer ses cartes marines.

Les effets sont extraordinaires. 50 millions de kilomètres en moins en 2011 pour les chauffeurs d'UPS, onze millions de litres de carburant économisés et plus de 27 tonnes d'émissions de dioxyde de carbone en moins, selon Jack Levis, directeur de la gestion des processus d'UPS. Le programme a également amélioré la sécurité et l'efficacité : l'algorithme compile des itinéraires avec moins de virages et moins de croisements aux carrefours, source fréquente d'accidents, de perte de temps et de consommation supplémentaire de carburant.

« La prévision nous a apporté des connaissances, déclare Levis d'UPS. Mais au-delà de la connaissance, il y a plus : la sagesse et la clairvoyance. Un jour, le système sera si intelligent qu'il prédira les problèmes et les réglera avant que l'utilisateur ait pris conscience que quelque chose n'allait pas. »

Au fil du temps, la géolocalisation et sa mise en données s'appliquent aussi massivement aux personnes. Durant des années, les opérateurs mobiles ont collecté et analysé des informations pour améliorer le niveau de service de leurs réseaux. Mais ces données sont de plus en plus souvent utilisées à d'autres fins et collectées par des tiers pour de nouveaux services. Certaines applications smartphone, par exemple, recueillent des informations relatives à la localisation même si l'application en elle-même ne dispose pas d'une fonctionnalité de géolocalisation. Dans d'autres cas, le but de l'application est de développer une entreprise fondée sur la connaissance des lieux que fréquentent ses utilisateurs. À titre d'exemple, le média social Foursquare, qui donne à ses utilisateurs la possibilité de signaler leur localisation (*check in*) et, par là, leurs lieux de sortie favoris. Il tire ses revenus des programmes de fidélisation, des recommandations de restaurants et d'autres services liés à la localisation que communiquent ses utilisateurs.

Cette collecte des données de localisation est en passe d'acquérir une valeur considérable. Elle donne la possibilité d'émettre des publicités ciblées vers l'individu-utilisateur, sur le lieu où il se trouve ou bien sur celui où il a prévu de se rendre. Les informations peuvent être agrégées pour faire éventuellement émerger des tendances. Par exemple, il devient possible aux entreprises de détecter des embouteillages sans avoir besoin de voir les voitures mais en amassant des données de localisation : le nombre et la vitesse des téléphones circulant sur une autoroute fournissent ces informations. La société AirSage produit ainsi des rapports sur la circulation en temps réel dans plus de 100 villes aux États-Unis en analysant quotidiennement 15 milliards de relevés de positionnement relatifs aux déplacements de millions d'abonnés au téléphone portable. Deux autres sociétés de géolocalisation, Sense Networks et Skyhook, peuvent dire dans quels quartiers la vie nocturne est la plus animée dans une ville, ou évaluer le nombre de manifestants présents lors d'une manifestation, en utilisant leurs données de localisation.

Mais ce sont les utilisations non commerciales de la géolocalisation qui pourraient se révéler les plus importantes de toutes. Alex Sandy Pentland, directeur du Human Dynamics Laboratory au MIT, et Nathan Eagle ont été les pionniers de ce qu'ils ont appelé le *reality mining* (extraction de la réalité). Il s'agit pour eux d'établir des inférences et de faire des prédictions sur le comportement humain à partir du traitement d'énormes quantités de données collectées à partir des téléphones portables. Une de leurs études leur a permis de détecter qui avait contracté la grippe avant qu'il (ou elle) ne le sache, grâce à l'analyse de ses déplacements et de ses habitudes d'appels. Dans le cas d'une épidémie de grippe mortelle, avec une telle capacité, il deviendrait possible d'informer à tout moment les responsables de la santé publique des zones les plus touchées et de sauver des millions de vies. Mais s'il se retrouvait entre des mains irresponsables, le *reality mining* pourrait avoir des conséquences terribles, comme nous le verrons plus loin.

Eagle, le fondateur de Jana, une start-up spécialisée dans le domaine des données sans fil, a utilisé les données agrégées de téléphones portables provenant de plus de 200 opérateurs de téléphonie mobile dans plus de 100 pays – quelque 3,5 milliards de gens répartis entre l'Amérique latine, l'Afrique et l'Europe – pour répondre à des questions qui intéressent particulièrement les directeurs de marketing, par exemple savoir combien de fois par semaine la lessive est faite dans un ménage. Mais il a aussi eu recours à des quantités massives de données pour examiner d'autres aspects comme la façon dont les villes atteignent la prospérité. En collaboration avec un collègue, il a combiné des données de localisation sur des abonnés de lignes téléphoniques mobiles prépayées en Afrique avec celles précisant le montant de la dépense au moment où ils rechargeaient leur ligne. Ce montant est fortement corrélé au revenu : les gens plus aisés achètent plus de minutes à la fois. Mais une des conclusions paradoxales d'Eagle est que les bidonvilles, loin de n'être que des foyers de pauvreté, jouent également le rôle de tremplins économiques. Le fait est que ces utilisations indirectes de données de localisation n'ont rien à voir avec l'acheminement de communications mobiles, but initial pour lequel les informations ont été recueillies. Une fois que la localisation est mise en données, de nouvelles utilisations potentielles émergent, accompagnées d'une nouvelle création de valeur.

Quand les interactions deviennent des données

Les nouveaux horizons de la mise en données nous concernent plus personnellement : elles visent nos relations, nos expériences et nos états d'âme. L'idée de mise en données est à la base de la création de nombreuses sociétés de réseaux sociaux sur Internet. Ces plateformes ne sont pas là simplement pour nous offrir un moyen de trouver des amis et collègues et de rester en contact avec eux, elles peuvent se saisir d'éléments intangibles de notre vie quotidienne et les transformer en données servant à de nouveaux usages. Facebook a mis en données nos relations ; elles ont toujours été là et constituaient des informations, mais elles n'avaient jamais été définies formellement comme des données jusqu'à l'arrivée du « graphe social » de Facebook. En créant un moyen facile pour tout un chacun d'enregistrer puis de partager ses réflexions (dont les traces se seraient auparavant effacées) avec d'autres gens, Twitter a permis la mise en données des sentiments. Tout comme Maury a transformé de vieux journaux de bord, LinkedIn a mis en données nos expériences professionnelles passées, et a transformé ces informations en prévisions : les personnes que nous pourrions connaître ou un poste que nous souhaiterions décrocher.

L'utilisation des données en est encore au stade embryonnaire. Facebook, pour sa part, a fait montre d'une patience avisée, sachant que ses utilisateurs s'affoleraient si l'entreprise dévoilait prématurément un trop grand nombre de ses nouvelles visées concernant leurs données. D'ailleurs, la société continue à adapter son modèle d'entreprise (et sa politique de confidentialité) au volume et au type de collecte de données qu'elle souhaite pratiquer. De ce fait, les critiques qu'elle essuie sont beaucoup plus axées sur les informations qu'elle est capable de collecter que sur l'usage effectif qu'elle fait de ces données. Facebook comptait plus d'un milliard d'utilisateurs en 2012, et plus de 100 milliards de liens entre « amis ». Cela représente à l'arrivée et en termes de graphe social 10 % du total de la population mondiale totale, mis en données et disponibles pour être communiqué à une seule société.

Les utilisations potentielles sont extraordinaires. Un certain nombre de start-up ont cherché à adapter le graphe social dans le but de l'utiliser comme indicateur de l'évaluation des cotes de crédit. Qui se ressemble s'assemble, telle est l'idée : les esprits prudents se lient d'amitié entre eux, les paniers percés itou. Si cela réussit, Facebook pourrait bien créer la prochaine agence d'évaluation du crédit. Les vastes ensembles de données

détenus par les médias sociaux pourraient bien être à la base de nouvelles entreprises dont l'activité se déploierait bien au-delà des fonctionnalités superficielles tels le partage de photos, les mises à jour de statuts et les « j'aime ».

Les données de Twitter ont également été utilisées d'intéressantes façons. On pourrait penser que les 400 millions de tweets envoyés quotidiennement en 2012 par plus de 140 millions d'utilisateurs ne sont pas grand-chose d'autre que du bla-bla sans queue ni tête. Et c'est souvent le cas. Il n'en demeure pas moins que la société effectue la mise en données de pensées, d'humeurs et d'interactions, toutes informations qui n'auraient jamais pu être recueillies auparavant. Twitter a conclu des accords avec deux entreprises, Data-Sift et Gnip, pour vendre l'accès à ses données. (Tous les tweets sont publics, mais l'accès à la source a un coût.) De nombreuses entreprises font de l'analyse de tweets, parfois à l'aide d'une technique appelée analyse de sentiments, afin de recueillir le retour informationnel de groupes de clients ou de juger l'impact de certaines campagnes.

Deux fonds de pension, Derwent Capital à Londres et London et MarketPsych en Californie, ont commencé à analyser le contenu de tweets mis en données et à les interpréter comme indicateurs de placements en Bourse. (Leurs stratégies réelles de négociation ont été tenues secrètes : plutôt que d'investir dans des entreprises bénéficiant de beaucoup de tapage, les deux sociétés ont peut-être misé contre ces dernières.) Elles vendent maintenant leurs informations aux traders. MarketPsych, quant à elle, s'est associée avec Thomson Reuters pour proposer pas moins de 18 864 indices boursiers distincts dans 119 pays, mis à jour toutes les minutes, sur des états émotionnels tels l'optimisme, la morosité, la joie, la peur, la colère et même des thèmes comme l'innovation, les litiges et les conflits. Les données sont utilisées plutôt par des ordinateurs que par des humains : les cracks en maths de Wall Street, connus sous le nom de *quants*, insèrent les données dans leurs modèles algorithmiques pour rechercher des corrélations invisibles susceptibles de se muer en profits. La fréquence elle-même à laquelle des tweets sont émis sur un sujet peut prédire diverses choses, comme les recettes de films hollywoodiens, selon Bernardo Huberman, un des pères de l'analyse des réseaux sociaux. Avec un collègue chez HP, il a élaboré un modèle qui suit le rythme de publication des nouveaux tweets. Ils ont ainsi pu prévoir le succès d'un film avec de

meilleurs résultats qu'au moyen d'autres variables prédictives habituellement utilisées.

Les possibilités ne s'arrêtent pas là. Si les messages Twitter sont limités à 140 caractères, les métadonnées – autrement dit, les « informations à propos des informations » – associées à chaque tweet sont tout aussi précieuses. Elles comprennent 33 éléments distincts. Certains ne semblent pas très utiles, comme le fond d'écran du twitto (c'est ainsi qu'on appelle l'utilisateur de Twitter) ou le logiciel qu'il emploie pour accéder au service. Mais d'autres métadonnées sont extrêmement intéressantes, tels la langue du twitto, sa géolocalisation et le nombre et les noms des personnes qu'il suit et de ceux qui le suivent. Dans une étude qui a fait l'objet d'un article dans *Science* en 2011, une analyse de 509 millions de tweets sur deux ans émis par 2,4 millions de personnes dans 84 pays a révélé que les humeurs des gens suivent des schémas quotidiens et hebdomadaires similaires dans toutes les cultures du monde – constatation impossible auparavant. Les humeurs ont été mises en données.

Dans la mise en données, il ne s'agit pas seulement de convertir en une forme analysable des attitudes et des sentiments, mais aussi le comportement humain, lequel est difficile à cerner, en particulier dans un contexte de communauté élargie et de sous-groupes qui la composent. Le biologiste Marcel Salathé de l'université de l'État de Pennsylvanie et l'ingénieur logiciel Shashank Khandelwal ont découvert en analysant des tweets que l'attitude des gens concernant la vaccination correspondait à la probabilité qu'ils se fassent effectivement vacciner eux-mêmes contre la grippe. Fait important, ils ont poussé leur investigation en utilisant les métadonnées sur l'identité de ceux qui étaient connectés parmi les « suiveurs » sur Twitter. Ils ont constaté qu'il pouvait exister des sous-groupes de personnes non vaccinées. Ce qui fait l'originalité de cette recherche c'est qu'à la différence d'autres études, telle Google Suivi de la grippe, qui s'est servie de données agrégées pour examiner l'état de santé des gens, l'analyse des sentiments effectuée par Salathé a vraiment prédit leur comportement en matière de santé.

Ces premiers résultats indiquent quelle sera à coup sûr la prochaine étape de la mise en données. À l'instar de Google, un ensemble de médias sociaux fonctionnant en réseau comme Facebook, Twitter, LinkedIn, Foursquare et d'autres sont assis sur une énorme malle au trésor, en l'occurrence des informations mises en données qui, une fois analysées,

permettront de mieux saisir la dynamique sociale à tous les niveaux, de l'individu à la société dans son ensemble.

La mise en données de tout

Avec un peu d'imagination, une foultitude de choses peut être transformée en données – et nous surprendre au passage. Dans le même esprit que ce que fait le professeur Koshimizu sur les arrière-trains à Tokyo, IBM a obtenu un brevet américain en 2012 portant sur la « Sécurisation des locaux à l'aide d'une technologie informatique de surface ». Cette formulation en langage d'avocat spécialisé en propriété intellectuelle désigne un revêtement de sol tactile, un genre d'écran de smartphone géant. Les utilisations potentielles sont innombrables : elle pourrait permettre d'identifier des objets placés dessus. Dans une version simple, elle serait capable d'éteindre les lumières d'une pièce ou ouvrir une porte quand une personne entre. Mais, le plus important, elle pourrait identifier des individus à leur poids ou à leur posture quand ils marchent ou se tiennent debout. Elle pourrait détecter si une personne fait une chute et ne se relève pas, fonction importante dans le cas de personnes âgées. Les commerçants pourraient connaître le flux de visiteurs dans leurs magasins. Une fois le sol mis en données, il n'y a plus de limite aux utilisations qu'il est possible d'en faire.

La mise en données d'un maximum de choses possible pourrait bien être réalisée plus tôt qu'on ne le pense. Voyez par exemple le mouvement « Quantified self^e » qui réunit un groupe disparate de passionnés de fitness, d'obsédés de l'univers médical et d'accros à la technologie qui mesurent chaque élément de leur corps et chaque aspect de leur vie dans le but de vivre mieux – ou du moins, de découvrir de nouveaux aspects d'une manière détaillée, chose qui leur aurait été impossible auparavant. Le nombre de pratiquants de l'« auto-mesure » est réduit pour l'instant, mais en passe d'augmenter.

Avec l'arrivée des smartphones et le faible coût de la technologie informatique, la mise en données des actes les plus essentiels de la vie n'a jamais été aussi facile. Une kyrielle de start-up donne la possibilité aux gens de mesurer leurs ondes cérébrales et ainsi de découvrir la structure de leur sommeil au cours de la nuit. Une société, Zeo, a déjà constitué la plus grande base de données sur le sommeil et trouvé des différences entre sexes en termes de durée de sommeil paradoxal. Asthmapolis a fixé un capteur à

un inhalateur pour le traitement de l'asthme pour suivre et localiser le patient par GPS ; en agrégeant les informations, la société est en mesure d'identifier les déclencheurs environnementaux des crises, comme la proximité de certaines cultures.

Les sociétés Fitbit et Jawbone proposent des capteurs pour mesurer l'activité physique et le sommeil. Une autre société, Basis, propose un bracelet qui surveille les paramètres vitaux de celui qui le porte, notamment le rythme cardiaque et la conductivité de la peau, tous deux mesures du stress. Obtenir les données est une opération plus facile et moins intrusive que jamais. En 2009, Apple a obtenu un brevet pour des oreillettes qui collectent des données sur le taux d'oxygénation du sang, le rythme cardiaque et la température corporelle.

Il y a beaucoup à apprendre de la mise en données du fonctionnement du corps. Des chercheurs de l'université de Gjøvik en Norvège et la société Drawi Biometrics ont mis au point une application pour smartphones qui analyse la démarche d'une personne et utilise les informations, recueillies par des capteurs intégrés, comme système de sécurité pour autoriser l'accès au téléphone. Entre-temps, deux professeurs du GeorgiaTech Research Institute, Robert Delano et Brian Parise, ont élaboré une autre application appelée iTrem qui utilise l'accéléromètre intégré au téléphone pour surveiller les tremblements du corps provoqués par la maladie de Parkinson et d'autres troubles neurologiques. Cette application est une vraie bénédiction tant pour les médecins que pour les patients. Ces derniers peuvent éviter des tests coûteux effectués au cabinet d'un spécialiste ; quant aux médecins, elle leur permet de suivre à distance le stade d'invalidité d'un patient et la qualité de réponse aux traitements. Selon des chercheurs de Kyoto, un smartphone est à peine un peu moins efficace pour mesurer les tremblements que l'accéléromètre tridimensionnel des équipements médicaux spécialisés ; il peut donc être utilisé en toute fiabilité. Là encore, un peu de désordre surpasse l'exactitude.

Dans la plupart des cas que nous venons de citer, les informations recueillies sont converties en données afin d'être réutilisées. Cela peut se faire quasiment partout et s'appliquer à presque tout. GreenGoose, une start-up de San Francisco, vend de minuscules capteurs de mouvement que l'on peut fixer sur des objets afin de savoir à quelle fréquence ils sont utilisés. En les plaçant sur un paquet de fil dentaire, un arrosoir ou une boîte de litière pour chat, il est possible de mettre en données l'hygiène dentaire,

l'entretien des plantes ou celui des animaux de compagnie. L'enthousiasme soulevé par l'« Internet des objets » – l'intégration de puces, de capteurs et de modules de communication dans des articles de la vie courante – relève en partie de la mise en réseau mais tout autant de la mise en données de tout notre environnement.

Quand le monde est mis en données, les limites des utilisations potentielles des informations s'arrêtent là où s'arrête l'ingéniosité de l'homme. Maury a mis en données les voyages de nombreux marins en procédant avec persévérance à une lourde opération de compilation manuelle et a ainsi ouvert la porte à une connaissance extraordinaire d'une formidable valeur. Aujourd'hui, il nous est possible d'effectuer des tâches similaires bien plus rapidement, à bien plus grande échelle et dans de très nombreux contextes, grâce à nos nouveaux outils (statistiques et algorithmes) et à un équipement indispensable (processeurs numériques et stockage). À l'ère des big data, même un postérieur a ses bons côtés.

À certains égards, nous nous retrouvons au beau milieu d'un vaste projet qui rivalise par son infrastructure avec celles du passé, des aqueducs romains à l'Encyclopédie du siècle des Lumières. Nous avons du mal à nous en rendre compte parce que le projet actuel est tellement nouveau, que nous sommes en plein dedans et parce que, contrairement à l'eau qui coule sur les aqueducs, le produit de nos travaux est intangible. Ce projet, c'est la mise en données. À l'instar de ces autres avancées, il provoquera des bouleversements fondamentaux dans notre société.

Les aqueducs ont contribué à la croissance des villes ; la presse à imprimer a contribué à l'avènement du siècle des Lumières et les journaux ont permis la naissance de l'État-nation. Ces infrastructures étaient axées sur les flux – celui de l'eau, celui du savoir. Tout comme le téléphone et l'Internet. La mise en données, elle, constitue un enrichissement essentiel de la compréhension de l'humain. Grâce aux big data, notre monde ne nous apparaît plus telle une suite d'événements à interpréter comme des phénomènes naturels ou sociaux, mais comme un univers composé essentiellement d'informations.

Depuis au moins un siècle, les physiciens suggèrent que c'est le cas – que ce ne sont pas les atomes mais les informations qui sont à la base de tout. Certes, cela peut paraître « ésotérique ». Mais dans de nombreux cas, grâce à la mise en données, nous pouvons agir sur notre existence dans ses

aspects physiques et intangibles qu'il nous est maintenant possible de capter, voire de calculer.

La vision du monde sous forme d'informations, comme un océan de données qui peut être exploré à une profondeur et avec une ampleur toujours plus grandes, nous montre la réalité sous une perspective que nous n'avions pas auparavant. C'est un nouvel état d'esprit qui pourrait percoler dans tous les domaines de la vie. Aujourd'hui, notre société s'est dotée de connaissances mathématiques parce que nous pensons que le monde est compréhensible grâce à elles. Et nous prenons pour acquis que le savoir peut se transmettre à travers le temps et l'espace parce que l'idée du mot écrit est profondément ancrée en nous. Demain, il pourrait bien y avoir une « prise de conscience des big data » chez les prochaines générations – l'hypothèse que dans toutes nos actions entre une composante quantitative, et que la société peut apprendre des choses de ces données qui lui sont indispensables. Il paraît sans doute nouveau à la plupart d'entre nous que la myriade de dimensions de la réalité puisse être transformée en données. Mais à l'avenir, nous considérerons sûrement cette mise en données comme un acquis.

À terme, l'impact de la mise en données pourrait éclipser celui des aqueducs et des journaux, et peut-être rivaliser avec l'imprimerie et Internet en nous donnant les moyens de cartographier le monde de manière quantifiable et analysable. Pour l'instant, les utilisateurs les plus avertis de la mise en données demeurent les entreprises, qui utilisent les big data pour créer de nouvelles formes de valeur – thème du prochain chapitre.

[a.](#) Depot of charts and instruments. Il fut renommé quelques années plus tard Naval Observatory (Observatoire naval).

[b.](#) Nom de l'United States Navy (Marine des États-Unis) durant la guerre de Sécession.

[c.](#) « Computus », en latin, signifie calcul. En français, le terme « ordinateur » vient du latin « ordinator » : celui qui met de l'ordre, l'ordonnateur.

[d.](#) Téléphone « espion ». (*N.d.É.*)

[e.](#) « La quantification de soi ».

6

Valeur

À la fin des années 1990, Internet est devenu assez rapidement un terrain incontrôlable, rebutant et peu convivial. Les messageries étaient inondées de spams et des « spambots » (robots délivreurs de spams) engorgeaient les forums en ligne. En 2000, Luis von Ahn, âgé de vingt-deux ans et fraîchement diplômé de l'université, imagina un moyen de résoudre le problème : obliger quiconque voulant s'inscrire sur un site Internet à prouver qu'il était un humain et non un robot. Il se mit donc en quête d'un système à la portée des utilisateurs mais difficile à utiliser par les machines.

Il eut l'idée de présenter des lettres déformées et difficiles à lire qui s'afficheraient lors de l'inscription. Là où la personne serait parfaitement capable de les déchiffrer et de saisir correctement le texte en quelques secondes, les ordinateurs seraient bloqués. Yahoo appliqua sa méthode et vit le fléau des spambots s'écrouler du jour au lendemain. Von Ahn nomma son invention Captcha (Completely Automated Public Turing Test to Tell Computers and Humans Apart^a). Cinq ans plus tard, des millions de captchas étaient saisis quotidiennement.

Captcha valut à von Ahn une immense célébrité et lui ouvrit les portes de l'université Carnegie Mellon pour y enseigner l'informatique après l'obtention de son doctorat. Elle contribua également à lui faire recevoir, à vingt-sept ans, le prestigieux prix « des génies » décerné par la Fondation Mac Arthur et doté d'une bourse de 500 000 dollars. Mais quand il se rendit compte qu'à cause de lui, des millions de gens se retrouvaient chaque jour à perdre du temps à saisir laborieusement des lettres déformées – des tonnes d'informations qui ne servaient plus à rien par la suite – il se trouva un peu moins génial !

Il trouva alors une nouvelle méthode qui allait utiliser de manière plus productive toute cette puissance de calcul humain, et la nomma fort à propos ReCaptcha. Au lieu de saisir des lettres aléatoires, les internautes devaient saisir deux mots issus de projets de scannage de textes qu'un logiciel de reconnaissance optique de caractères n'était pas parvenu à déchiffrer. Dans ce système, l'un des deux mots est destiné à confirmer ce que les autres utilisateurs ont saisi, ce qui indique qu'il s'agit d'un humain ; le second est un mot nouveau qui nécessite une clarification. Par souci d'exactitude, un même mot flou doit être saisi correctement par cinq personnes différentes en moyenne avant que le système ne soit sûr qu'il s'agit du bon mot. Ces données avaient deux fonctions : la première était de prouver que l'utilisateur était humain, la seconde était de parvenir à déchiffrer les mots flous figurant dans des textes numérisés.

La valeur ajoutée de ce système est immense, si l'on pense à ce qu'il en coûterait de confier le travail à des êtres humains. À raison de 10 secondes environ par intervention, au rythme actuel de 200 millions de ReCaptchas par jour, cela représente au total un demi-million d'heures quotidiennement. En 2012, le salaire minimum aux États-Unis était de 7,25 dollars. S'il avait fallu payer des gens pour clarifier des mots que l'ordinateur ne parvient pas à déchiffrer, il aurait fallu déboursier 4 millions de dollars par jour, soit plus d'un milliard de dollars par an. Alors que von Ahn a conçu un système pour effectivement faire faire le travail, mais gratuitement. Vu l'intérêt majeur de cette technologie, Google l'acheta à von Ahn en 2009 pour ensuite en proposer l'utilisation gracieusement à tous les sites Web intéressés ; aujourd'hui, elle est intégrée sur près de 200 000 sites, notamment Facebook, Twitter et Craigslist.

L'histoire de ReCaptcha souligne l'importance de la réutilisation des données. Avec les big data, la valeur des données change. À l'ère du numérique, les données qui ont perdu leur rôle dans les transactions se sont souvent transformées en une marchandise en soi. Dans un monde de big data, il s'agit d'une mutation nouvelle. La valeur des données se déplace de leur utilisation initiale vers ses futures utilisations potentielles, un phénomène qui a un impact sérieux, notamment sur la manière dont les entreprises évaluent les données en leur possession et aussi sur l'autorisation d'accès qu'elles y donnent. Il permet aux entreprises de changer de modèle, et les y contraint parfois. Il modifie la perception que les organisations ont des données et de leur façon de les utiliser.

Les informations ont toujours joué un rôle essentiel dans les transactions commerciales. Les données permettent par exemple la détermination des prix, indicateur de la quantité à produire. Une dimension des données parfaitement comprise. Certains types d'informations sont depuis longtemps négociés sur les marchés : contenu de livres, d'articles, de musique et de films, par exemple, tout comme les informations financières, le cours des actions notamment. Sont venues s'y ajouter ces dernières décennies les données personnelles. Aux États-Unis, les courtiers spécialisés en données, comme Acxiom, Experian et Equifax, demandent de coquettes sommes pour fournir des dossiers complets d'informations personnelles sur des centaines de millions de consommateurs. Avec Facebook, Twitter, LinkedIn et d'autres plateformes de médias sociaux, nos relations, nos opinions, nos goûts personnels et notre style de vie quotidienne ont rejoint l'ensemble d'informations personnelles déjà disponibles à notre sujet.

En résumé, même si les données présentent depuis longtemps un intérêt, elles ont été considérées comme accessoires en comparaison des activités de base de l'entreprise, ou alors limitées à des catégories relativement restreintes comme la propriété intellectuelle ou les informations personnelles. En revanche, à l'ère des big data, *toutes* les données seront considérées comme précieuses par définition.

Quand nous disons « toutes les données », nous entendons par là même les bribes d'information les plus simples, les plus anodines en apparence. Il n'y a qu'à penser aux relevés d'un détecteur de chaleur placé sur une machine-outil ; ou au flux en temps réel de coordonnées GPS, aux relevés d'accéléromètres et au niveau de carburant d'un véhicule de livraison (ou d'un parc de 60 000 véhicules !) ; ou alors aux milliards de requêtes ou au prix de quasiment chaque siège sur tous les vols commerciaux aux États-Unis depuis plusieurs années.

Jusqu'à récemment, comme il était difficile de collecter, de stocker et d'analyser ces types de données, on trouvait tout aussi difficile d'en extraire la valeur potentielle. Si l'on reprend l'exemple célèbre de la manufacture d'épingles cité par Adam Smith, qui illustre ainsi son propos sur la division du travail au XVIII^e siècle, il aurait fallu que des observateurs soient mobilisés en permanence, jour après jour, pour surveiller tous les ouvriers, en notant le détail de leurs mesures et de la production avec une plume d'oie. À l'époque où les économistes classiques étudiaient les facteurs de

production (terre, travail et capital), l'idée d'exploiter les données n'existait pratiquement pas. Le coût de leur collecte puis de leur utilisation est resté relativement élevé jusqu'à récemment, même s'il a diminué au cours des deux derniers siècles.

Ce qui fait la différence, c'est qu'aujourd'hui, une bonne partie des contraintes inhérentes à la collecte a disparu. Grâce à de nouvelles technologies, de vastes quantités d'informations peuvent être recueillies puis enregistrées à un coût marginal. Les données relatives aux personnes sont fréquemment collectées de manière passive, sans effort ou presque de ces dernières, voire à leur insu. Et avec la forte baisse du coût de stockage, le choix de les conserver au lieu de les éliminer se justifie plus facilement. Pour toutes ces raisons, les données sont plus que jamais disponibles, en plus grand nombre à un prix moindre. Depuis cinquante ans, le stockage numérique a vu son prix baisser de moitié à peu près tous les deux ans tandis que la capacité de stockage, elle, a été multipliée par 50 millions. Si l'on prend l'exemple d'entreprises dont l'activité est axée sur les informations comme Farecast ou Google – où les faits bruts entrent par un bout de la chaîne de production numérique et des informations traitées ressortent à l'autre bout –, on peut considérer que les données correspondent à une nouvelle ressource ou à un facteur de production.

La valeur immédiate de la plupart des données est évidente pour ceux qui les collectent. Il est probable en fait qu'ils les recueillent en ayant à l'esprit un objectif précis. Les magasins s'assurent d'une bonne comptabilisation financière en collectant des données sur les ventes. Les usines s'assurent de leur conformité aux normes de qualité en contrôlant leur production. Les sites Web analysent et optimisent le contenu qu'ils présentent à leurs visiteurs en enregistrant leur moindre clic – voire parfois les mouvements du curseur de la souris. Ces premières utilisations ont justifié la collecte et le traitement des données. Lorsque Amazon enregistre non seulement les achats de livres par les clients mais aussi les pages Web qu'ils ont simplement visitées, cette entreprise sait qu'elle utilisera les données pour faire des recommandations personnalisées. De la même manière, Facebook suit les « mises à jour de statuts » et les « j'aime » des utilisateurs de façon à déterminer les publicités les plus pertinentes à afficher sur son site Web et à en tirer un revenu.

Contrairement aux choses matérielles – la nourriture que nous consommons, une bougie qui brûle – la valeur des données ne s'use pas

quand on s'en sert ; elles peuvent être traitées et re-traitées. Les informations sont ce que les économistes appellent un bien « non rival » : la consommation de ce bien par l'un n'empêche pas l'autre de le consommer. Il est donc possible à Amazon, quand elle fait des recommandations à ses clients, d'utiliser des données issues d'anciennes transactions – et de recommencer la même opération à plusieurs reprises, non seulement auprès du client qui en est la source mais aussi auprès de nombreux autres.

Tout comme elles peuvent servir plusieurs fois le même but, les données peuvent aussi être exploitées à de multiples fins. C'est un aspect important à comprendre si l'on veut déterminer le volume d'informations qui nous sera utile à l'ère des big data. Nous avons vu qu'une partie de ce potentiel s'est déjà concrétisée quand, par exemple, en fouillant dans sa base de données de vieux reçus de ventes, Walmart a repéré cette corrélation juteuse entre l'ouragan et les ventes de biscuits fourrés Pop-Tarts.

Tout cela indique que la vraie valeur des données dépasse largement celle qui a été extraite lors de leur première utilisation. Cela signifie également que, même si elles ont déjà fait une première utilisation de leurs données ou même quelques autres d'infime valeur, les entreprises peuvent efficacement les exploiter du moment quelles les réutilisent plusieurs fois.

La « valeur d'option » des données

Réutiliser des données pour en tirer toute la valeur. Si l'on veut vraiment saisir ce que cela signifie, les voitures électriques en sont une bonne illustration. C'est à un éventail assez ahurissant d'aspects logistiques, tous en rapport avec la vie des batteries, qu'elles vont devoir leur succès ou leur échec comme mode de transport. Il faut que les automobilistes puissent recharger les batteries rapidement et simplement, sans que la consommation d'énergie ne déstabilise le réseau, ce à quoi doivent veiller les compagnies d'électricité. Si le réseau de stations-service dont nous disposons aujourd'hui est efficace, nous ne connaissons pas encore les besoins en termes de recharge des véhicules électriques ni combien il faudra de stations de recharge, ni où.

Chose frappante, le problème ne se pose pas tant en termes d'infrastructures qu'en termes d'information – les big data jouant un rôle important dans la solution. Lors d'un essai effectué en 2012, IBM, en

collaboration avec la société californienne Pacific Gas and Electric Company ainsi que le constructeur automobile Honda, ont collecté d'énormes quantités d'informations dans le but de répondre à certaines questions essentielles : quand et où les voitures électriques devront-elles effectuer une recharge et qu'est-ce que cela implique pour le système d'alimentation électrique. IBM a conçu un modèle prédictif très élaboré fondé sur plusieurs informations d'entrée : le niveau de la batterie du véhicule, le lieu où se trouve le véhicule, le moment de la journée et les bornes disponibles dans les stations de recharge proches. La société a combiné ces données avec la consommation d'électricité instantanée sur le réseau ainsi qu'avec les antécédents de consommation électrique. À partir de l'analyse des énormes flux de données en temps réel et passés en provenance de multiples sources, IBM a ainsi pu déterminer les meilleures périodes pour recharger les batteries de véhicules électriques. Elle a également pu indiquer les meilleurs emplacements pour construire des stations de recharge. À terme, le système devra prendre en compte les différences de prix entre les stations de recharge proches en fonction également des prévisions météo : exemple, s'il fait beau et qu'une station proche alimentée à l'énergie solaire regorge d'électricité, ou, à l'inverse si la météo annonce une semaine de pluie durant laquelle ses panneaux solaires seront à l'arrêt.

Le système analyse des informations obtenues dans tel ou tel but et les réutilise à une autre fin – en d'autres termes, les données passent d'une utilisation première à une utilisation secondaire. Avec le temps, leur valeur augmente. Dans la voiture, l'indicateur de charge indique au conducteur quand recharger la batterie. La compagnie d'électricité collecte les données concernant l'utilisation du réseau afin de gérer sa stabilité. Ce sont les utilisations premières. Les deux ensembles de données peuvent avoir des utilisations secondaires et se forger une nouvelle valeur en permettant de déterminer quand et où effectuer une recharge et où construire des stations-service pour véhicules électriques. De surcroît, d'autres informations sont également intégrées, comme la localisation du véhicule et la consommation passée du réseau. Le traitement des données par IBM ne se fait pas qu'une fois, mais de multiples fois, lors d'une continuelle mise à jour du tableau de consommation d'énergie des voitures électriques ainsi que de la sollicitation du réseau électrique.

Pour la véritable valeur des données, on peut reprendre la métaphore de l'iceberg : seule la pointe est visible, et tout le reste immergé. Les sociétés innovantes qui l'ont compris vont extraire ce potentiel caché et en tirer des bénéfices peut-être gigantesques. En clair, la vraie valeur des données ne peut se mesurer qu'en prenant en compte tous les emplois qu'on pourra en faire à l'avenir, outre l'utilisation actuelle. Nous en avons déjà cité de nombreux exemples. C'est en exploitant les données concernant des billets d'avion vendus par le passé que Farecast a su prédire les tarifs aériens. Google a pu révéler la prévalence de la grippe en réutilisant les requêtes des internautes. En compilant les journaux de bord tenus par d'anciens capitaines Maury a pu indiquer le tracé des courants océaniques.

L'importance de cette réutilisation des données n'est pas encore, il faut le dire, appréciée à sa juste valeur dans le monde des affaires ni par la société. Il y avait peu de cadres de Con Edison à New York qui auraient pu imaginer l'utilisation à des fins de prévention des accidents d'informations à propos des câbles, notés dans des registres d'entretien vieux d'un siècle. Il a fallu une nouvelle génération de statisticiens et une nouvelle vague de méthodes et d'outils pour que se révèle la valeur des données. Ce n'est que récemment que de nombreuses sociétés Internet et entreprises technologiques elles-mêmes en ont pris conscience.

Peut-être est-il utile de voir les données comme les physiciens le font pour l'énergie. Ils parlent d'énergie « stockée » ou « potentielle », qui existe à l'état latent mais qui demeure inexploitée. Prenez un ressort comprimé ou un ballon posé au sommet d'une colline. L'énergie contenue dans ces objets demeure latente – potentielle – jusqu'à sa libération, disons, quand le ressort est relâché ou que le ballon roule le long de la pente. L'énergie de ces objets est alors devenue « cinétique » parce qu'ils se sont mis en mouvement et peuvent exercer une force sur d'autres objets. Après leur utilisation initiale, les données conservent donc une valeur latente ; leur potentiel est stocké à la manière du ressort ou du ballon, jusqu'à ce que l'utilisation secondaire en libère la puissance. À l'ère des big data, nous disposons enfin de l'état d'esprit, de l'ingéniosité et des outils pour exploiter la valeur cachée des données.

In fine, leur valeur réside dans les résultats finals, obtenus en les utilisant de toutes les manières possibles. Une liste apparemment infinie dans laquelle il faudra prendre des options, c'est-à-dire faire des choix. La somme de ces choix donnera leur importance aux données : leur « valeur

d'option ». Alors que jadis, nous étions prêts à supprimer les données une fois utilisées, à l'ère des big data, on se rend compte qu'il s'agit peu ou prou d'une véritable mine de diamants, qui demeure productive de façon quasi magique bien après que sa valeur essentielle a été exploitée. Trois moyens efficaces font apparaître cette valeur d'option : la réutilisation des bases des données ; la fusion de plusieurs ensembles de données ; la recherche d'ensembles de données qui font « d'une pierre deux coups ».

La réutilisation des données

Les requêtes constituent un exemple classique permettant une réutilisation innovante de données. Au premier abord, les informations paraissent sans intérêt une fois qu'elles ont été utilisées à leur première fin. Ainsi, après une courte interaction entre un consommateur et un moteur de recherche, une liste de sites Web et de publicités s'est affichée, qui remplissait une fonction précise à cet instant donné. Mais d'anciennes requêtes peuvent se révéler extraordinairement précieuses. Par exemple pour connaître les goûts des consommateurs. C'est le genre d'information qu'une société comme Hitwise, spécialisée dans la mesure du trafic Web et appartenant au courtier en données Experian, veut donner à ses clients. Les responsables du marketing peuvent ainsi faire appel à Hitwise pour savoir si la couleur rose sera à la mode ce printemps ou si le noir fait son retour. Google met gratuitement à la disposition du public une version de son outil d'analyse de trafic. La société a lancé en partenariat avec BBVA, la deuxième plus grande banque espagnole, un service de prévisions économiques en matière de tourisme et vend des indicateurs économiques en temps réel fondés sur les recherches d'internautes. Pour savoir si les prix de l'immobilier grimpent ou chutent, la Bank of England utilise des requêtes relatives aux logements.

Les entreprises qui n'ont pas mesuré l'importance de la réutilisation de données ont reçu une belle leçon. Ainsi, à ses débuts, Amazon a signé un accord avec AOL pour utiliser la technologie qui sous-tend le site d'e-commerce d'AOL. Pour la plupart des gens, cela ressemblait à un accord d'externalisation ordinaire. Mais ce qui intéressait vraiment Amazon, explique Andreas Weigend, l'ancien directeur scientifique de la société, c'était de disposer de données concernant ce que les utilisateurs d'AOL allaient regarder sur Internet et ce qu'ils y achetaient, dans l'idée

d'améliorer la performance de son moteur de recommandation. Malheureux AOL à qui ce point échappa et qui ne vit la valeur de ses données qu'en termes d'utilisation initiale – les ventes. Ceux d'Amazon, plus malins, savaient les bénéfices à tirer d'une utilisation secondaire des données. Prenons aussi le cas de l'entrée de Google sur le terrain de la reconnaissance vocale avec GOOG-411, un service téléphonique d'annuaire professionnel, une expérience menée de 2007 à 2010. Tout géant qu'il était, Google ne disposait pas de sa propre technologie de reconnaissance vocale ; il lui fallait donc en trouver une exploitable en sous-licence. Un accord fut conclu avec Nuance, leader dans ce domaine, ravi d'avoir décroché un client aussi convoité. Mais à l'époque Nuance était nulle en matière de larges volumes de données : le contrat ne stipulait pas qui devait conserver les traductions vocales, et c'est Google qui garda les enregistrements. L'analyse de données permet d'évaluer la probabilité qu'une bribe de voix numérisée corresponde à un mot spécifique, chose essentielle pour améliorer la technologie de la reconnaissance vocale, voire créer un nouveau service. À l'époque, Nuance considérait que son activité portait sur la concession de licences de logiciel, et non sur l'analyse de données. Sitôt après avoir reconnu son erreur, la société s'est demandé comment recueillir des données. Elle a alors commencé à passer des accords pour l'utilisation de son service de reconnaissance vocale avec des opérateurs de téléphonie mobile et des fabricants de téléphones portables.

La bonne nouvelle, c'est que la réutilisation des données peut produire de la valeur pour des organisations qui font peu usage actuellement de vastes ensembles de données mais les collectent ou les contrôlent, à l'instar des entreprises classiques dont l'activité s'exerce principalement hors ligne. Il est bien possible qu'elles soient assises sur des mines d'informations inexploitées. Certaines sociétés peuvent avoir collecté des données, les avoir utilisées une fois (ou même pas du tout), et les avoir conservées en raison tout simplement du faible coût de stockage – dans des sortes de « cimetières de données », comme les spécialistes scientifiques de données appellent ces lieux où sont enfouies des informations qui datent.

Bien sûr, les sociétés Internet et les entreprises de haute technologie, par leur présence en ligne, sont en bonne place pour exploiter le déluge de données dont elles font la collecte et elles ont une longueur d'avance sur le reste du secteur en termes d'analyse de ces données. Il demeure que toutes les entreprises auraient à y gagner. Chez McKinsey & Company, les

consultants mentionnent ainsi une entreprise de logistique, dont ils taisent le nom, qui a remarqué qu'avec le processus de livraison des marchandises remontait une foule d'informations sur les expéditions de produits à travers le monde. Flairant une bonne affaire, la société a créé un département spécial pour la vente de données agrégées sous forme de prévisions commerciales et économiques. En d'autres termes, elle a créé une version hors ligne du service requêtes anciennes de Google. Prenez le cas de SWIFT, le système interbancaire mondial utilisé pour les virements électroniques. La société a constaté une corrélation entre les paiements et l'activité économique mondiale. Ainsi, SWIFT propose des prévisions de croissance du PIB en se basant sur ce que révèlent les transferts de fonds transitant sur son réseau.

Certaines entreprises, grâce à leur position dans la chaîne de valeur de l'information, ont la capacité de collecter d'énormes quantités de données, même si elles n'en ont pas un besoin immédiat ou si elles ne sont pas en mesure de les réutiliser. Les opérateurs de téléphonie mobile, parce qu'ils acheminent les appels vers leurs abonnés, collectent ainsi des informations sur les lieux où ces derniers se trouvent. Pour ces sociétés, ces données n'ont qu'un usage technique restreint. Mais elles prennent de la valeur quand elles sont réutilisées par d'autres, qui diffusent de la publicité et font des promotions personnalisées et géolocalisées. Il arrive aussi que la valeur réside non pas dans des points de données individuels mais dans ce qu'ils révèlent pris tous ensemble. Ainsi, les entreprises de géolocalisation comme AirSage et Sense Networks évoquées au chapitre précédent peuvent vendre des informations indiquant les lieux où les gens se rassemblent le vendredi soir ou bien précisant le rythme auquel progressent des véhicules dans les bouchons. Ces masses d'informations peuvent aussi servir à chiffrer la valeur de biens immobiliers ou les tarifs de panneaux publicitaires.

Bien utilisées, même les informations les plus banales peuvent acquérir une valeur particulière. Revenons à l'exemple des opérateurs de téléphonie mobile : ils conservent les données sur le lieu et l'instant où les téléphones se connectent aux antennes-relais, et sur l'intensité du signal. Depuis longtemps, ils s'en servent pour parfaire la performance de leurs réseaux, décider où étoffer leur infrastructure, où la mettre à niveau. Mais ces données ont bien d'autres utilisations potentielles. Elles pourraient permettre aux fabricants de téléphones mobiles de mieux savoir ce qui influence l'intensité du signal, par exemple, et ainsi voir comment

améliorer la qualité de réception de leurs appareillages. Les opérateurs mobiles sont réticents à monnayer ces informations de crainte d'enfreindre le règlement sur la protection des données personnelles. Mais ils commencent à assouplir leur position du fait qu'ils sont en difficulté financière et voient dans leurs données une source potentielle de revenu. En 2012, Telefonica, important opérateur espagnol de téléphonie mobile implanté dans plusieurs pays, a même créé une société indépendante, Telefonica Digital Insights, de vente de données anonymes et agrégées de localisation d'abonnés auprès des détaillants et d'autres sociétés.

Données recombinaisons

Il arrive aussi que la valeur latente se révèle encore autrement : en combinant un ensemble de données avec un autre, probablement très différent. De nouveaux résultats émergent en mélangeant les données selon de nouvelles procédures. Un bon exemple est celui de l'enquête publiée en 2011, destinée à déterminer si les téléphones cellulaires augmentaient ou non le risque de cancer – question cruciale quand on sait qu'il y a près de 6 milliards de téléphones portables dans le monde, soit quasiment un par personne sur la planète. De nombreuses études ont échoué auparavant. Soit elles faisaient appel à de trop petits échantillons, soit les périodes examinées étaient trop courtes ou alors elles se fondaient sur des données autodéclarées et truffées d'erreurs. Une équipe de chercheurs de la Danish Cancer Society a cependant eu l'idée d'adopter une nouvelle méthode en se fondant sur des données recueillies antérieurement.

Depuis l'arrivée des téléphones mobiles sur le marché danois, les données concernant l'ensemble des abonnés provenaient des opérateurs de téléphonie mobile. Ce que l'enquête a ciblé, c'étaient les possesseurs de portable de 1987 à 1995, à l'exception des abonnés via leur entreprise ainsi que d'autres types d'abonnés aux données socioéconomiques indisponibles. Soit un ensemble de 358 403 personnes. Par ailleurs, il faut savoir que l'État danois a tenu un registre national de tous les patients atteints de cancer, parmi lesquels 10 729 personnes atteintes de tumeurs du système nerveux central, durant la période 1990-2007, la période de suivi. Enfin, il existe au Danemark un registre national qui mentionne le niveau d'études et le revenu disponible de chaque habitant. Pour leur enquête, les chercheurs ont combiné les trois ensembles de données, puis vérifié si les utilisateurs

de téléphones mobiles présentaient des taux plus élevés de cancer que les non-abonnés. Ils se sont également posé la question de savoir si, parmi les abonnés, le risque de développer un cancer était plus grand chez ceux qui possédaient un téléphone cellulaire depuis plus longtemps.

En dépit de la taille de l'enquête, les données n'étaient ni déstructurées ni imprécises : elles devaient en effet répondre à des normes de qualité exigeantes à des fins médicales, commerciales ou démographiques. Le mode de collecte ne laissait place à aucun biais lié au sujet de l'enquête. De fait, ces données avaient été obtenues des années plus tôt, sans aucun rapport avec cette recherche. Point le plus important, l'enquête ne s'est pas fondée sur un échantillon, mais sur un volume de données proche de $N = \text{tous}$: la quasi-totalité des cas de cancer, et presque tous les utilisateurs de portables, soit 3,8 millions d'années-personnes propriétaires d'un portable. De surcroît, parce que les données incluaient quasiment tous les cas, les chercheurs ont pu tenir compte de sous-populations, telles les personnes au niveau de revenu élevé.

Finalement, aucune augmentation du risque de cancer associé à l'utilisation de téléphones portables n'a été mise en évidence. D'où un silence quasi total dans les médias lors de la publication de l'étude en octobre 2011 dans le *British Medical Journal*. Si un lien avait été découvert, en revanche, l'enquête aurait fait la une des journaux dans le monde entier, et la méthodologie des « données recombinaisons » aurait été saluée.

Avec les big data, le tout vaut plus que la somme des parties, et quand nous recombinaisons les sommes de plusieurs ensembles de données, cette somme vaut à son tour plus que chacune d'elles. Aujourd'hui, les utilisateurs d'Internet connaissent bien les *mashups*^b de base, qui combinent deux sources de données ou plus, de façon nouvelle. Par exemple, Zillow, le site Web dédié à l'immobilier, superpose informations et prix de l'immobilier sur le plan d'un quartier aux États-Unis. Il analyse par ailleurs une pile de données, comme les récentes transactions réalisées dans le quartier et les spécifications des biens immobiliers, ce qui lui permet de prédire la valeur de telle ou telle maison aux caractéristiques spécifiques, dans une zone particulière. La présentation visuelle rend les données plus accessibles. Mais avec les big data, nous pouvons aller bien au-delà. L'enquête danoise sur le cancer nous a donné un bon aperçu de ces possibilités.

Données extensibles

Il est utile de prévoir dès le départ comment rendre les données « extensibles », pour pouvoir les adapter et les réutiliser à de multiples usages. Ce n'est pas toujours possible – du fait que les idées sur des utilisations potentielles peuvent venir bien après la collecte des données – mais il existe des moyens de susciter des utilisations multiples d'un même ensemble de données. Exemple, les caméras de surveillance. Certains commerçants les installent dans leur magasin pour repérer les voleurs à l'étalage mais pas seulement. Cela leur permet aussi de suivre le flux des clients qui circulent dans les allées et de voir devant quels rayons ils s'arrêtent pour regarder. Cette seconde fonction leur permet de concevoir un meilleur agencement du magasin et d'évaluer l'efficacité des campagnes de marketing. Auparavant, ces caméras vidéo n'étaient là que pour la sécurité. Maintenant, elles sont considérées comme un investissement susceptible d'augmenter le chiffre d'affaires.

Comme on pouvait s'y attendre, Google est l'un des meilleurs en matière d'intégration de la notion d'extensibilité. Les véhicules de son programme Street View, qui ont suscité une certaine controverse en prenant des photos de maisons et de rues dans de multiples villes, ont aussi englouti des données GPS, vérifié des informations cartographiques et même récupéré des noms de réseaux wifi (et, sans doute de manière illégale, le contenu qui circulait sur les réseaux sans fil ouverts). En un seul circuit effectué par ces voitures, une myriade de flux de données distincts à tout moment a été amassée. Là où les données ont pris leur caractère extensible, c'est quand Google a commencé à s'en servir pour de multiples utilisations secondaires en plus de leur utilisation initiale. Par exemple, les données GPS ont permis d'améliorer le service de cartographie de l'entreprise, des données indispensables au fonctionnement des voitures sans conducteur utilisées pour le programme Street View.

Le coût supplémentaire qu'entraîne la collecte de flux multiples ou de points de données beaucoup plus nombreux dans chaque flux, est souvent marginal. Il paraît donc évident qu'il faut recueillir autant de données que possible, et les rendre extensibles en envisageant dès le départ les utilisations secondaires potentielles. Cela accroît leur valeur d'option. Le principe est de chercher à « faire d'une pierre deux coups » – en effectuant la collecte selon un certain processus qui permette à l'ensemble de données

recueillies d'être réutilisé dans de multiples cas. Ainsi, les données font double usage.

Dépréciation de la valeur des données

La baisse du coût de stockage des données numériques incite les entreprises à conserver des données qu'elles pensent réutiliser (aux mêmes fins ou à des fins similaires). Mais il existe une limite à leur utilité.

Par exemple, quand des entreprises comme Netflix et Amazon convertissent les achats des clients, leur navigation sur le site et les comptes rendus de livres en recommandations, elles peuvent être tentées d'utiliser les mêmes documents à de multiples reprises sur plusieurs années. On pourrait alors avancer que, tant qu'une entreprise n'est pas freinée par des lois et des réglementations comme les lois sur la protection de la vie privée, elle ne devrait pas s'abstenir d'utiliser des documents numériques indéfiniment, ou du moins aussi longtemps que c'est possible sur un plan économique. Dans la réalité, ce n'est pas si simple.

La plupart des données perdent en effet de leur intérêt avec le temps. Dans ce cas, continuer à se fier à de vieilles données non seulement n'apporte pas de valeur ajoutée mais annihile celle de données plus récentes. Prenons l'exemple d'un livre que vous avez acheté il y a dix ans chez Amazon, et qui ne correspondrait plus à vos centres d'intérêt actuels. Si Amazon se base sur la trace de cet achat remontant à dix ans en arrière pour vous recommander d'autres livres, il y a peu de chances que vous achetiez ces titres – ou même que vous vous intéressiez aux recommandations que vous ferait le site ultérieurement. Comme Amazon base ses recommandations à la fois sur des informations périmées et sur des plus récentes, encore valables, la présence des anciennes diminue la valeur des nouvelles.

Par conséquent, la société a fortement intérêt à limiter l'utilisation des données à la période où elles sont encore productives. Elle doit procéder en permanence à un nettoyage de ses mines de données et éliminer les informations qui ont perdu de la valeur. La difficulté est de savoir lesquelles sont devenues inutiles. Fonder sa décision simplement sur la période est rarement approprié. Amazon et d'autres sociétés ont donc mis au point des modèles élaborés qui leur permettent de faire le tri entre données utiles et non pertinentes. Par exemple, si un client regarde un livre recommandé sur

la base d'un achat antérieur ou l'achète, on peut en déduire que l'achat ancien est encore représentatif des goûts actuels du client. Ainsi, on peut évaluer le degré d'utilité des données plus anciennes, et modéliser avec plus de précision le « taux de dépréciation » des informations.

Toutes les données ne perdent pas de leur valeur au même rythme ni de la même façon. Cela explique pourquoi certaines sociétés sont persuadées qu'il leur faut conserver les données le plus longtemps possible, même si les organismes de réglementation ou le public souhaitent les voir éliminées ou rendues anonymes au bout d'un certain temps. Par exemple, Google s'est longtemps refusé à supprimer les adresses IP complètes des utilisateurs correspondant à des requêtes anciennes. (La société n'efface que les derniers chiffres au bout de neuf mois pour rendre la requête partiellement anonyme. Ainsi, elle a encore les moyens de comparer les données d'une année sur l'autre, par exemple pour les recherches liées aux achats de périodes de fêtes – mais uniquement au niveau régional, sans affiner jusqu'au consommateur individuel.) Connaître la localisation des visiteurs du site contribue aussi à améliorer la pertinence des résultats. Par exemple, si beaucoup d'internautes vivant à New York font une recherche sur la Turquie – et cliquent sur les liens de sites en rapport avec le pays –, l'algorithme fera aussi remonter les pages de ces sites dans le classement affiché pour d'autres New-Yorkais. Même si la valeur des données baisse pour atteindre certains buts en vue desquels on les utilise, leur valeur d'option peut demeurer forte.

La valeur des traces numériques

La réutilisation des données prend parfois une forme astucieuse et cachée. Les entreprises du Web qui recueillent des données sur toutes les interactions des utilisateurs peuvent ensuite traiter chacune d'elles comme un signal distinct à utiliser comme retour d'information pour personnaliser le site, améliorer un service ou créer un produit numérique entièrement nouveau. Nous en voyons une illustration intéressante dans l'histoire de deux systèmes de correction orthographique.

Au cours des vingt dernières années, Microsoft a mis au point un correcteur orthographique très efficace pour son logiciel Word. L'entreprise a procédé en comparant un dictionnaire des termes correctement orthographiés fréquemment mis à jour, avec la succession de lettres saisies

par les utilisateurs. Le dictionnaire a déterminé les termes connus ; les variantes proches ne figurant pas dedans ont été traitées par le système comme des fautes de frappe à corriger. La compilation et la mise à jour du dictionnaire exigeant beaucoup d'efforts, le correcteur orthographique de Microsoft Word n'est disponible que pour les langues les plus courantes. La création et la maintenance de cet outil coûtent des millions à Microsoft.

Prenons maintenant le cas de Google. Le moteur est doté du correcteur orthographique sans doute le plus achevé, dans pratiquement toutes les langues. Le système s'améliore et intègre en permanence de nouveaux termes – conséquence indirecte de l'utilisation quotidienne du moteur de recherche par les internautes. Une erreur de saisie sur le terme « iPad » ? Google l'a prévue. Sur « Obamacare » ? Google la repère.

De surcroît, son correcteur orthographique n'a vraisemblablement rien coûté à Google, qui l'a mis au point en réutilisant les fautes d'orthographe dans les termes saisis sur son moteur de recherche, au sein des trois milliards de requêtes qu'il gère quotidiennement. Une boucle rétroactive intelligente informe le système du terme que les utilisateurs voulaient en réalité saisir. Parfois, ceux-ci « donnent » explicitement à Google la réponse à la question posée par le moteur en haut de la page des résultats – « Voulez-vous dire *épidémiologie* ? » – en cliquant dessus pour commencer une nouvelle recherche avec le terme correct. Ou bien la page Web finalement visitée va signaler implicitement la bonne orthographe, probablement du fait que le terme saisi est plus fortement corrélé avec le terme correctement orthographié qu'avec celui incorrectement orthographié. (Cet aspect est plus important qu'il n'y paraît : comme le correcteur orthographique ne cesse de s'améliorer, les utilisateurs ne se soucient même plus de saisir correctement leurs requêtes, sachant que Google peut les traiter sans difficulté de toute façon.)

Le système de correction orthographique de Google apporte la démonstration que des données « de mauvaise qualité », « incorrectes » ou « défectueuses » ont tout de même une utilité. Fait intéressant, Google n'a pas été le premier à avoir cette idée. Vers l'an 2000, Yahoo a entrevu la possibilité de créer un correcteur orthographique à partir des fautes d'orthographe de ses utilisateurs. Mais l'idée est restée sans suite. Les données relatives à d'anciennes requêtes étaient globalement considérées comme du rebut. De même, Infoseek et Alta Vista, moteurs de recherche populaires il y a quelques années, possédaient chacun la base de données de

termes mal orthographiés la plus complète au monde pour l'époque, mais ils n'en avaient pas mesuré la valeur. Leur système, par un procédé invisible aux yeux des utilisateurs, traitait les fautes de frappe comme des « termes apparentés » et effectuait une recherche dessus. Mais celle-ci était fondée sur des dictionnaires qui indiquaient explicitement au système ce qui était correct, et non sur la somme vivante et dynamique des interactions des utilisateurs.

Google est le seul à avoir reconnu que le rebut était de l'or qui, ramassé et fondu, pouvait se transformer en beau lingot brillant. Selon l'un des ingénieurs de haut niveau de Google, les performances de son correcteur orthographique sont supérieures d'au moins un ordre de grandeur à celles du correcteur de Microsoft – mais quand il lui a été demandé de donner des précisions sur cette affirmation, il a admis ne pas avoir effectué de mesures fiables pour la corroborer. Et il a vivement rejeté l'idée que le développement de l'outil avait été « gratuit ». Même si la matière première – les fautes d'orthographe – n'a rien coûté, Google a sans doute dépensé beaucoup plus que Microsoft pour mettre au point le système, a-t-il affirmé avec un large sourire.

La différence de démarche des deux sociétés est extrêmement révélatrice. Pour Microsoft, la correction orthographique n'avait de valeur que dans un seul but, le traitement de texte. Google, pour sa part, en a saisi l'utilité plus profonde. La société a non seulement utilisé les fautes de frappe pour élaborer le meilleur correcteur orthographique et le plus à jour au monde dont la fonction est d'améliorer les recherches sur Internet, mais elle a appliqué ce système à de nombreux autres services, comme la fonction « saisie automatique » lors des recherches, dans Gmail, dans Google Docs, et même dans son système de traduction.

Un terme spécialisé décrit maintenant la trace que laissent dans leur sillage tous ceux qui utilisent des outils numériques : « traces numériques ». Il s'agit d'un sous-produit constitué par les données provenant des actions et mouvements des individus à travers le monde. Dans le contexte d'Internet, le terme correspond aux interactions en ligne des utilisateurs : sur quoi ils cliquent, combien de temps ils restent à lire ou regarder une page, ce que survole le curseur de la souris, le texte qu'ils saisissent, et plus encore. De nombreuses sociétés conçoivent leurs systèmes de manière à recueillir ces traces numériques et à les recycler, afin d'améliorer un service existant ou d'en créer de nouveaux. Google est le leader incontesté dans ce

domaine. Il applique à nombre de ses services le principe selon lequel il faut « tirer des enseignements des données » de manière récursive. Chaque action exécutée par l'utilisateur est considérée comme un signal à analyser et à intégrer dans le système.

Par exemple, Google connaît parfaitement le nombre de fois où quelqu'un cherche un terme ainsi que des termes associés, et le nombre de fois où il a cliqué sur un lien puis est revenu à la page de recherche parce que la réponse trouvée ne le satisfaisait pas, avant de reprendre sa recherche. Il sait s'il a cliqué sur le premier lien de la huitième page – ou s'il a complètement abandonné la recherche. La société n'a sans doute pas été la première à avoir cette idée, mais elle l'a mise en pratique avec une efficacité extraordinaire.

Ces informations sont très précieuses. Si de nombreux utilisateurs ont tendance à cliquer sur un résultat de requête qui se trouve en bas de la page de résultats, cela indique qu'il est plus pertinent que ceux placés au-dessus, et l'algorithme de classement de Google sait où le remonter automatiquement sur la page en vue de recherches ultérieures. (Cela fonctionne de même pour les publicités.) « Nous aimons les enseignements que nous donnent de grands ensembles de données imprécises », lance un membre de l'équipe de Google.

À la base de nombreux services comme la reconnaissance vocale, les filtres anti-spam, la traduction, et bien plus encore, on trouve les traces numériques. Ainsi, quand les utilisateurs indiquent à un programme de reconnaissance vocale qu'il a mal compris ce qu'ils ont dit, en fait, ils permettent au système de « s'entraîner » et donc de s'améliorer.

De nombreuses entreprises commencent à concevoir leur propre système de collecte et d'utilisation des informations sur ce mode. Dans les débuts de Facebook, Jeff Hammerbacher, le premier « expert scientifique en données » (et parmi les premiers à être crédités d'avoir inventé le terme), a examiné l'abondant trésor de traces numériques qu'avait amassé la société. Avec son équipe, il a découvert une variable prédictive importante : elle indiquait que quelqu'un allait interagir sur le site s'il avait vu l'un de ses amis faire la même chose (publier du contenu, cliquer sur une icône, etc.). Aussi Facebook a-t-il modifié son système pour rendre plus visibles les activités des amis, ce qui a déclenché un cercle vertueux de nouvelles contributions au site.

L'idée se répand, bien au-delà du secteur d'Internet, dans les sociétés qui disposent d'un moyen pour recueillir les retours utilisateurs. Les liseuses de livres numériques, par exemple, recueillent des quantités massives de données sur les goûts littéraires et les habitudes de leurs utilisateurs : le temps qu'ils mettent à lire une page ou quelques chapitres, le lieu où ils lisent, s'ils tournent la page après l'avoir à peine parcourue ou s'ils ferment le livre une fois pour toutes. Chaque fois que les utilisateurs soulignent un passage ou bien mettent une annotation dans la marge, ces appareils enregistrent l'interaction. La capacité de rassembler ce type d'informations transforme la lecture, qui fut longtemps un acte solitaire, en une sorte d'expérience communautaire.

Une fois agrégées, les traces numériques peuvent fournir aux éditeurs et aux auteurs des informations auxquelles ils n'auraient jamais eu accès auparavant de manière quantifiable : les goûts, les aversions et les habitudes de lecture des gens. Ces informations sont précieuses au plan commercial. On peut tout à fait imaginer que des sociétés de livres numériques les vendent à des éditeurs pour améliorer le contenu et la structure des livres. Par exemple, l'analyse de données de Barnes & Noble à partir de sa liseuse Nook a révélé que les lecteurs avaient tendance à abandonner à mi-parcours les ouvrages de non-fiction volumineux. Cette découverte a donné à la société l'idée de créer une série intitulée « NookSnaps » : des œuvres courtes sur des thèmes actuels comme la santé et les questions d'actualité.

Prenons un autre exemple, celui des programmes d'enseignement universitaire en ligne tels Udacity, Coursera et edX. Ils suivent les interactions des étudiants sur Internet pour voir ce qui fonctionne le mieux sur le plan pédagogique. Le nombre d'étudiants par classe peut atteindre des dizaines de milliers, ce qui fournit des quantités extraordinaires de données. Les enseignants sont maintenant à même de voir si un fort pourcentage d'étudiants a visionné à nouveau une partie du cours, signe qu'un point particulier n'était pas clair pour eux. En donnant un cours Coursera sur l'apprentissage machine, Andrew Ng, professeur à Stanford, a remarqué que près de 2 000 étudiants avaient mal compris une question particulière dans leurs devoirs – mais qu'ils donnaient exactement la même réponse incorrecte. Visiblement, ils faisaient la même erreur, mais laquelle ?

Après enquête, il a découvert qu'ils inversaient deux équations algébriques dans un algorithme. Maintenant, quand d'autres étudiants commettent la même erreur, le système ne se contente pas de leur signaler

que c'est faux ; il leur suggère de vérifier leurs calculs mathématiques. Le système applique également la méthode des big data : il analyse chaque commentaire sur le forum qui a été lu par les étudiants et vérifie s'ils ont terminé correctement leur devoir, ceci dans le but de prédire la probabilité qu'un étudiant qui a lu un commentaire donné fournisse des résultats corrects, de façon à déterminer quels sont les commentaires que les étudiants ont le plus intérêt à lire. Ces éléments d'information, absolument inaccessibles auparavant, pourraient changer l'enseignement et l'apprentissage à tout jamais.

Les traces numériques constituent un avantage concurrentiel énorme pour les sociétés. Elles pourraient aussi devenir une puissante barrière à l'entrée pour les concurrents. Voici pourquoi : si une société récemment créée concevait un site de commerce électronique, un réseau social ou un moteur de recherche nettement supérieur aux leaders actuels – Amazon, Google ou Facebook –, elle aurait du mal à rivaliser avec eux non seulement pour des raisons d'économie d'échelle, d'effets de réseau ou de marque mais aussi parce qu'une bonne partie de la performance de ces entreprises leaders est due aux traces numériques qu'elles collectent à partir des interactions des clients et qu'elles réintègrent dans le service qu'elles offrent. Un nouveau site Web d'enseignement en ligne pourrait-il avoir le savoir-faire suffisant pour rivaliser avec un concurrent en possession d'une quantité gigantesque de données qui lui servent à savoir ce qui donne les meilleurs résultats ?

La valeur des données ouvertes

Aujourd'hui, nous sommes tentés de penser que des sites comme Google et Amazon sont les pionniers des big data. Or, ce sont bien sûr les gouvernements qui ont, les premiers, recueilli des informations à grande échelle, et qui continuent à surpasser n'importe quelle entreprise privée vu l'énorme volume de données qu'ils contrôlent. La différence avec les détenteurs de données du secteur privé est que les gouvernements peuvent dans bien des cas obliger les citoyens à leur fournir des informations, au lieu de procéder par persuasion ou de proposer une contrepartie. Les gouvernements continueront donc à amasser de vastes richesses en termes de données.

Les leçons des big data s'appliquent autant au secteur public qu'aux entités commerciales : la valeur des données publiques est latente et requiert une analyse novatrice pour la faire émerger. Mais en dépit de leur position privilégiée pour obtenir des informations, les gouvernements ont souvent été inefficaces à les utiliser. Récemment, une idée a pris de l'importance, à savoir que le meilleur moyen d'exploiter la valeur des données publiques était d'en donner accès au secteur privé et à la société dans son ensemble. Un principe sous-tend également cette idée : quand l'État recueille des données, il le fait au nom des citoyens, et la société devrait donc y avoir accès (excepté dans un nombre limité de cas, notamment si cela pouvait menacer la sécurité nationale ou porter atteinte à la vie privée d'autrui).

Cette idée a conduit à d'innombrables initiatives de « libération des données publiques » à travers le monde. Faisant valoir que les gouvernements ne sont que les dépositaires des informations qu'ils collectent, et que le secteur privé et la société seraient plus innovants, les partisans des « données ouvertes » demandent aux instances officielles de divulguer publiquement les données, à des fins civiques et commerciales. Pour que cela soit opérationnel, naturellement, les données doivent se présenter sous une forme normalisée, dans un format lisible par la machine pour être traitées rapidement. Autrement, les informations n'auraient de public que le nom.

L'idée de données publiques ouvertes a bénéficié d'un sérieux coup de pouce du président Barack Obama quand, à l'occasion de sa prise de fonctions le 21 janvier 2009, il a émis un mémorandum présidentiel ordonnant aux chefs des différents organismes fédéraux de publier le plus de données possible. « Dans le doute, c'est l'ouverture qui prévaut », a-t-il ordonné. C'était une déclaration remarquable, surtout comparée à l'attitude de son prédécesseur, qui avait donné ordre aux organismes de faire exactement le contraire. L'ordre d'Obama a entraîné la création d'un site Web intitulé data.gov, un répertoire d'informations librement accessibles émanant du gouvernement fédéral. Le site est passé de 47 ensembles de données en 2009 à près de 450 000 provenant de 172 organismes au moment de son troisième anniversaire en juillet 2012.

Même en Grande-Bretagne, pays pourtant réticent à cette ouverture des données, où beaucoup d'informations gouvernementales étaient protégées par le droit d'auteur de la Couronne et où la procédure de licence d'exploitation avait été difficile et coûteuse à mettre sur pied (comme les

codes postaux des entreprises de commerce électronique), des progrès considérables ont été enregistrés. Le gouvernement britannique a édicté des règles pour encourager l'ouverture des données et a soutenu la création de l'Open Data Institute^c codirigé par Tim Berners-Lee, l'inventeur du World Wide Web, pour promouvoir de nouvelles utilisations des données ouvertes et les moyens de les libérer de l'emprise de l'État.

L'Union européenne a également annoncé des initiatives d'ouverture des données qui pourraient à court terme s'élargir à l'ensemble du continent. Dans d'autres parties du monde, des pays comme l'Australie, le Brésil, le Chili et le Kenya ont établi et mis en œuvre des stratégies de données ouvertes. À travers le monde, un nombre croissant de villes et de municipalités dans certains pays ont, elles aussi, adopté des pratiques de données ouvertes, comme l'ont fait des organisations internationales telles que la Banque mondiale, qui a publié des centaines d'ensembles de données relatives à des indicateurs économiques et sociaux qui étaient auparavant en accès restreint.

En parallèle, des communautés de développeurs du Web et de penseurs visionnaires se sont constituées autour de ces données pour réfléchir au moyen d'en tirer le meilleur parti, comme Code for America et Sunlight Foundation aux États-Unis, et Open Knowledge Foundation en Grande-Bretagne.

Un des premiers exemples des possibilités offertes par les données ouvertes concerne un site Web du nom de FlyOnTime.us. Les visiteurs du site peuvent découvrir de manière interactive (parmi de nombreuses autres corrélations) quelles sont les probabilités qu'une intempérie provoque un retard dans les vols à un aéroport donné. Le site Web combine des informations sur les vols et sur la météorologie issues de sources de données officielles en libre accès sur Internet. Il a été créé par des défenseurs de l'ouverture des données qui voulaient démontrer l'utilité des informations amassées par le gouvernement fédéral. Le code logiciel du site est lui-même un code source ouvert, pour que d'autres puissent en tirer des enseignements et le réutiliser.

FlyOnTime.us laisse parler les données, et elles racontent souvent des choses surprenantes. On peut voir que pour les vols allant de Boston à l'aéroport de LaGuardia à New York, les voyageurs doivent s'attendre à des retards en cas de brouillard deux fois plus longs qu'en cas de chute de neige. Ce n'est probablement pas ce que la plupart des gens auraient

imaginé tandis qu'ils arpentaient la salle d'embarquement ; c'est plutôt la neige qui aurait semblé fournir un motif de retard plus sérieux. Mais c'est le type d'information que les big data peuvent donner et elle est le résultat du traitement combiné de gros volumes de données de diverses origines : les données historiques sur les retards de vols issues du Bureau of Transportation Statistics^d, les informations actuelles sur les aéroports fournies par la Federal Aviation Administration^e, ainsi que d'anciens bulletins météorologiques provenant de la National Oceanic and Atmospheric Administration (NOAA)^f ainsi que les conditions météorologiques en temps réel émanant du National Weather Service^g. FlyOnTime.us démontre comment une entité qui ne collecte ni ne contrôle des flux d'informations, contrairement à ce que fait un moteur de recherche ou un grand détaillant, peut néanmoins produire de la valeur après avoir obtenu et traité des données.

Évaluer ce qui n'a pas de prix

Que les données soient ouvertes au public ou bien enfermées dans les coffres-forts des entreprises, leur valeur est difficile à mesurer. Prenons l'exemple des événements du vendredi 18 mai 2012. Ce jour-là, le jeune fondateur de Facebook Mark Zuckerberg (vingt-huit ans) a symboliquement sonné la cloche annonçant l'ouverture du NASDAQ (le plus grand marché électronique d'actions du monde) à partir du siège de la société à Menlo Park, en Californie. Le réseau social le plus grand du monde – qui se vantait de compter comme membres près d'un habitant de la planète sur dix – a entamé sa nouvelle vie de société cotée en Bourse. Le titre a immédiatement progressé de 11 %, comme le font les actions de nombreuses sociétés de nouvelles technologies le premier jour de négociation du titre. Mais, par la suite, quelque chose de bizarre s'est produit. Les actions de Facebook ont commencé à chuter. Les ordinateurs du NASDAQ ont eu un problème technique qui a temporairement interrompu la cotation du titre, ce qui n'a rien arrangé. Mais un problème plus grave se profilait. Pressentant que quelque chose n'allait pas, les placeurs des actions, dirigés par la banque d'investissement Morgan Stanley, ont littéralement soutenu la cotation pour qu'elle reste au-dessus du prix d'émission.

La veille, les banques de Facebook avaient fixé le prix d'introduction de l'action de la société à 38 dollars, soit une valorisation de 104 milliards de dollars. (À titre de comparaison, cela correspondait approximativement aux capitalisations boursières de Boeing, General Motors et Dell Computers réunies.) Quelle était la valeur effective de Facebook ? Dans ces comptes certifiés pour l'exercice 2011, utilisés par les investisseurs pour évaluer la société, Facebook avait déclaré 6,3 milliards de dollars d'actifs. Ce montant représentait la valeur de son matériel informatique, de son équipement de bureau et d'autres biens matériels. Quid de la valeur comptable des vastes stocks d'informations détenus par Facebook dans son coffre-fort ? Pratiquement zéro. Elle n'était pas incluse, même si la société n'est constituée quasiment de *rien* d'autre que des données.

La situation était vraiment paradoxale. Doug Laney, vice-président de la recherche chez Gartner, un bureau d'études de marché, analysa les chiffres de la période précédant l'offre publique initiale (OPI) et calcula que Facebook avait collecté 2,1 milliards d'éléments de « contenu monétisable » entre 2009 et 2011, notamment les « j'aime », les éléments publiés et les commentaires. Comparé à sa valorisation lors de l'OPI, cela signifie que chaque article, considéré comme un point de données distinct, avait une valeur de cinq centimes de dollar environ. On pourrait également dire que chaque utilisateur de Facebook valait près de 100 dollars, puisque les utilisateurs sont la source des informations que Facebook collecte.

Comment expliquer l'énorme divergence entre la valeur de Facebook selon les normes comptables (6,3 milliards de dollars) et sa valeur du marché initiale (104 milliards de dollars) ? Il n'y a aucun moyen d'y parvenir. Il est communément admis que la méthode actuelle pour déterminer la valeur d'une entreprise, qui consiste à regarder sa « valeur comptable » (c'est-à-dire essentiellement la valeur de sa trésorerie et de son actif corporel), n'est plus pertinente. En fait, l'écart entre la valeur comptable et la « valeur de marché » – le prix que la société atteindrait en Bourse ou si elle était carrément rachetée – se creuse depuis des décennies. Le Sénat américain a d'ailleurs tenu en 2000 des audiences sur la modernisation des règles en matière d'information financière, mises en place dans les années 1930, à une époque où il n'existait quasiment pas d'entreprises ayant une activité fondée sur les données. Cette question n'a pas de répercussions que sur le bilan d'une société : l'incapacité à évaluer

correctement la valeur d'une entreprise lui fait indubitablement courir des risques et entraîne une volatilité de ses actions.

La différence entre la valeur comptable d'une société et sa valeur de marché est considérée comme des « actifs incorporels ». Elle est passée de près de 40 % de la valeur des sociétés cotées en Bourse aux États-Unis au milieu des années 1980 aux trois quarts de leur valeur à l'aube du nouveau millénaire ; c'est un écart important ! Parmi les éléments considérés comme actifs incorporels figurent la marque, les talents et la stratégie – tout ce qui n'est pas matériel et ne relève pas du système financier et comptable formel. Mais, de plus en plus, on compte aussi dans les actifs incorporels les données appartenant aux sociétés et dont elles font usage.

En définitive, ce que cela indique, c'est qu'il n'y a actuellement aucun moyen évident d'évaluer les données. Le jour de la première cotation de Facebook, l'écart entre ses actifs formels et sa valeur incorporelle non comptabilisée approchait les 100 milliards de dollars. C'est parfaitement ridicule ! Cependant, ce type d'écart doit disparaître et cela arrivera le jour où les sociétés auront trouvé le moyen de comptabiliser la valeur de leurs données au bilan.

On s'oriente à petits pas vers cette direction. Un dirigeant de l'un des plus grands exploitants de réseau sans fil américains a confié que sa société reconnaissait l'immense valeur de ses données et se demandait s'il ne fallait pas les traiter comme un élément d'actif de la société en termes comptables formels. Mais dès que les avocats de la société ont eu vent de ce projet, ils l'ont arrêté net. Intégrer les données dans les livres comptables en rendrait probablement la société juridiquement responsable, ont-ils avancé. Et en hommes de loi vigilants, ils ont trouvé que c'était une fausse bonne idée.

Dans le même temps, des investisseurs se sont mis à tenir compte de la valeur d'option des données. Le niveau des actions de sociétés détentrices de données ou capables d'en collecter facilement aura des chances de monter, tandis que d'autres sociétés moins privilégiées en termes de données pourront voir leur valeur boursière diminuer très sensiblement. Ce phénomène ne nécessite pas de comptabiliser formellement les données au bilan. Les marchés et les investisseurs prendront en compte ces actifs incorporels dans leurs évaluations – malgré la difficulté de l'opération, comme l'atteste l'évolution en dents de scie du cours de l'action Facebook durant ses premiers mois de cotation. Mais avec la résolution des dilemmes d'ordre comptable et des questions de responsabilité, il est quasiment

certain que la valeur des données apparaîtra au bilan des entreprises et s'affirmera comme une nouvelle catégorie d'actif.

Comment les données seront-elles évaluées ? Il ne s'agira plus, pour calculer leur valeur, de simplement rajouter les profits issus de leur utilisation initiale. Mais procéder à leur estimation n'est pas évident, avec une valeur latente qui ne sera générée que par des utilisations secondaires encore inconnues. Cela fait penser aux difficultés rencontrées dans le domaine de la valorisation des produits dérivés avant le développement de l'équation de Black-Scholes dans les années 1970, ou bien dans l'évaluation des brevets, domaine où enchères, échanges, ventes privées, octroi de licences et multiples litiges créent lentement un marché de la connaissance. Du moins, la capacité à chiffrer la valeur d'option des données constitue-t-elle assurément une formidable opportunité pour le secteur financier.

Un bon point de départ serait de se pencher sur les différentes stratégies employées par les détenteurs de données pour en extraire la valeur. L'utilisation par une société de ses propres données est la plus évidente. Mais il est peu probable qu'elle soit capable de révéler toute la valeur latente de ses données. Une option plus ambitieuse serait donc d'octroyer à des tiers une licence d'utilisation. À l'ère des big data, de nombreux détenteurs de données pourraient envisager d'opter pour un accord aux termes duquel ils recevraient un pourcentage de la valeur extraite des données au lieu d'un forfait, à la manière dont les auteurs et interprètes perçoivent un pourcentage sur les ventes de livres, de musiques ou de films sous forme de redevances. Le même principe est valable pour la propriété intellectuelle dans le domaine de la biotechnologie : les concédants de licences peuvent exiger des redevances sur toute invention ultérieure issue de leur technologie. Ainsi, toutes les parties seront incitées à maximiser la valeur tirée de la réutilisation de données.

Cependant, il arrive que le preneur de licence ne soit pas en mesure d'extraire toute la valeur d'option ; c'est pourquoi les détenteurs de données peuvent ne pas souhaiter accorder l'exclusivité de leurs mines de données et préférer diversifier les preneurs, un choix qui pourrait devenir la norme et une façon également pour les détenteurs de données de ne pas mettre tous leurs œufs dans le même panier.

Un certain nombre de places financières sont apparues, qui testent des méthodes pour chiffrer la valeur des données. DataMarket, créée en Islande

en 2008, fournit l'accès à des ensembles de données libres provenant d'autres sources, comme les Nations unies, la Banque mondiale et Eurostat, et tire des revenus de la revente de données issues de prestataires commerciaux comme les sociétés d'études de marché. D'autres start-up ont tenté de jouer le rôle d'intermédiaires, de plateformes où des tiers cèdent leurs données à titre gratuit ou moyennant paiement. L'idée est de permettre à n'importe qui de vendre les données qu'il détient dans ses bases de données, tout comme eBay a fourni une plateforme au public pour vendre des objets personnels. Import.io encourage les entreprises à accorder des licences d'utilisation de leurs données qui seraient, sinon, extraites de leur site Web et utilisées gratuitement. Et Factual, la société créée par un ancien membre de l'équipe de Google, Gil Elbaz, met à disposition des ensembles de données qu'elle prend le temps de compiler elle-même.

La société Microsoft est entrée dans l'arène avec Windows Azure Marketplace, marché en ligne pour les données haut de gamme où les offres extérieures sont contrôlées, à la manière dont Apple supervise les offres sur sa plateforme de téléchargement d'applications App Store. Comme le conçoit Microsoft, une directrice de marketing travaillant sur une feuille de calcul Excel peut vouloir recouper les données internes à la société avec des prévisions de croissance du PIB provenant d'un cabinet de consultants en économie. Elle achètera alors d'un simple clic les données de différentes provenances, et celles-ci seront instantanément importées dans les colonnes de son tableau affiché à l'écran.

On ne peut dire pour l'instant comment les modèles d'évaluation évolueront. Mais une chose est certaine, c'est que les économies commencent à se constituer autour des données – et que de nouveaux acteurs en tireront des bénéfices, tandis qu'un certain nombre d'anciens seront probablement surpris de trouver un second souffle grâce à elles. Pour reprendre les propos de Tim O'Reilly, éditeur d'ouvrages d'informatique et sommité de la Silicon Valley, les données sont une plateforme, car elles constituent un élément essentiel à la création de nouveaux produits et modèles d'entreprise. La valeur fondamentale des données réside dans leur potentiel apparemment illimité de réutilisation : leur valeur d'option. La collecte d'informations est essentielle mais pas suffisante, du fait que la plus grande partie de la valeur des données se trouve dans leur utilisation, et non dans leur simple possession. Dans le chapitre suivant, nous

examinerons les modes effectifs d'utilisation des données ainsi que les entreprises émergentes sur le marché des big data.

- [a.](#) Test de Turing public entièrement automatique pour distinguer les ordinateurs des humains.
- [b.](#) Application composite de données.
- [c.](#) Institut des données ouvertes.
- [d.](#) Bureau des statistiques du transport.
- [e.](#) Agence gouvernementale de l'aviation civile.
- [f.](#) Agence nationale responsable de l'étude de l'océan et de l'atmosphère.
- [g.](#) Service météorologique des États-Unis.

7

Implications

En 2011, une ingénieuse start-up de Seattle, Decide.com, ouvre ses portes sur Internet avec un objectif incroyablement ambitieux : lancer un moteur de recherche capable de prévoir le prix d'une foultitude d'objets de consommation. Elle prévoit un démarrage relativement modeste, en se cantonnant au marché des produits high-tech, des téléphones mobiles aux télévisions à écran plat en passant par les appareils photo numériques. Ses ordinateurs absorbent alors les flux de données en provenance de sites d'e-commerce et ratissent le Web à la recherche de toute autre information sur les prix et les produits qu'ils peuvent y trouver.

En permanence et tout au long de la journée, les prix affichés sur Internet changent, grâce à une mise à jour dynamique fondée sur une infinité de facteurs complexes ; cela implique que la société doit collecter des informations concernant les prix à tout moment. Sont concernés non seulement de gros volumes de données mais aussi de gros volumes de textes car le système doit faire analyser les mots afin de distinguer un produit abandonné d'un autre à la veille d'être lancé sur le marché, informations qui doivent être communiquées aux consommateurs et qui influent sur les prix.

Un an plus tard, Decide.com analyse quatre millions de produits en utilisant plus de 25 milliards de prix observés. Le moteur de recherche est devenu capable de détecter dans la distribution des anomalies que personne n'avait jamais pu « voir » auparavant, par exemple que les prix de modèles anciens peuvent momentanément augmenter une fois que de nouveaux modèles sont mis sur le marché. La plupart des gens achètent l'ancien modèle en pensant qu'il sera moins cher, mais selon le moment où ils cliquent sur le bouton « acheter », ils peuvent tout à fait se retrouver à le

payer plus cher. Du fait que les magasins en ligne ont de plus en plus recours à des systèmes de tarification automatique, Decide.com peut repérer des flambées anormales de prix fixés par un procédé de tarification algorithmique et prévenir les consommateurs d'attendre. Selon les chiffres fournis par la société après mesure interne, les prévisions sont justes 77 % du temps et elles font potentiellement économiser près de 100 dollars par produit en moyenne aux acheteurs. Decide.com a une telle confiance dans ses informations qu'en cas de prédiction incorrecte, elle rembourse la différence de prix à ses membres (service payant).

En apparence, Decide.com n'était pas très différente de nombreuses start-up prometteuses qui visent à exploiter autrement des informations et à gagner honnêtement quelques dollars en contrepartie de leurs efforts. La spécificité de Decide.com ne réside pas dans les données : la société utilise des informations obtenues sous licence de sites d'e-commerce et autres, récupérées gracieusement sur le Web. Elle ne réside pas non plus dans l'expertise technique : la société ne fait rien qui soit d'une complexité telle que seuls les ingénieurs qui travaillent pour son compte maîtrisent la chose. Même si la collecte de données et les compétences techniques sont des facteurs importants, la spécificité de Decide.com se situe au niveau de la « tournure d'esprit » : la société est « orientée big data ». Elle a guetté l'opportunité et constaté que l'exploitation de certaines données pouvait révéler de précieux secrets. Et si Decide.com rappelle quelque peu Farecast (le site de prévisions de tarifs aériens), il y a une bonne raison à cela : toutes deux sont des créations d'Oren Etzioni.

Dans le chapitre précédent, nous avons expliqué que les données se transforment maintenant en une nouvelle source de valeur, phénomène dû en grande partie à leur valeur d'option, comme nous l'avons appelée, quand elles sont réutilisées à des fins nouvelles. L'accent a été mis sur les entreprises qui collectent des données. Notre attention va maintenant se tourner vers les sociétés utilisatrices de données, et la place qu'elles occupent dans la chaîne de valeur des informations. Nous examinerons l'importance de ce phénomène pour des organisations et pour des particuliers, en ce qui concerne à la fois leur carrière et leur vie quotidienne.

Trois types de sociétés, selon nous, ont émergé sur le marché des big data, qui se différencient par la valeur qu'elles offrent ; les données, les compétences et les idées.

D'abord, les données. Ces sociétés sont détentrices de données ou du moins elles y ont accès sans que ce soit forcément leur vocation d'origine. Ou alors elles ne possèdent pas les compétences adéquates ni l'esprit inventif nécessaire pour en extraire la valeur. Le meilleur exemple est Twitter, qui dispose visiblement d'un flux massif de données transitant dans ses serveurs, mais qui s'est tournée vers deux sociétés indépendantes pour octroyer à des tiers une licence d'utilisation.

Ensuite, les compétences. Il s'agit souvent de cabinets conseils, de fournisseurs de technologie et prestataires de solutions analytiques qui possèdent une expertise particulière et peuvent faire le travail, mais ne disposent sans doute pas des données elles-mêmes ni de l'ingéniosité nécessaire leur permettant de découvrir les utilisations les plus novatrices. Dans le cas de Walmart et les Pop-Tarts, par exemple, le distributeur s'est adressé aux spécialistes de Teradata, société d'analyse de données, pour aider à dégager de nouvelles idées.

En troisième lieu, la tournure d'esprit big data. Pour certaines entreprises, ce ne sont pas les données et le savoir-faire qui font principalement leur succès. Ce qui les distingue, c'est que leurs fondateurs et leurs employés ont des idées originales qui leur permettent de créer de nouvelles formes de valeur en exploitant autrement les données. Pete Warden en est un bon exemple : ce crac en informatique est le cofondateur de Jetpac qui émet des recommandations de voyages en se fondant sur les photographies téléchargées par les utilisateurs sur le site.

Jusqu'à aujourd'hui, ce sont les deux premiers points qui ont reçu le plus d'attention : les compétences, toujours rares, et les données, abondantes en apparence. Une nouvelle profession est alors apparue ces dernières années, l'« expert scientifique en données » (*data scientist*), qui associe les compétences du statisticien, du programmeur de logiciels, de l'infographiste et du conteur. Au lieu de s'user les yeux sur un microscope pour résoudre un des mystères de notre monde, l'expert scientifique en données scrute des bases de données pour faire ses découvertes. L'institut McKinsey (l'un des plus grands cabinets conseil mondiaux) profère de sombres prédictions sur la pénurie actuelle et future d'experts de ce type (lesquels aujourd'hui se plaisent à les citer pour se valoriser et gonfler leurs salaires).

Hal Varian, économiste en chef de Google, emploie cette formule célèbre pour qualifier le statisticien : le métier le plus « sexy » au monde.

« Si vous voulez réussir, vous devez être complémentaire et rare par rapport à un produit omniprésent et peu coûteux, déclare-t-il. Les données sont si largement disponibles et si importantes au plan stratégique que ce qui est rare, c'est le savoir nécessaire pour en extraire les enseignements. C'est pourquoi les statisticiens, les gestionnaires de bases de données et les professionnels de l'apprentissage vont vraiment se trouver dans une position formidable. »

L'accent mis sur les compétences et la minimisation de l'importance des données pourrait toutefois ne pas durer. Avec l'évolution du secteur, la généralisation des compétences vantées par Varian devrait pallier le manque de personnel qualifié. Qui plus est, on croit à tort que du simple fait de leur abondance, les données sont utilisables gratuitement ou qu'elles ont peu de valeur. En réalité, ce sont elles l'élément essentiel. Pour le comprendre, examinons les différents maillons de la chaîne de valeur des big data, et leur possible changement au fil du temps. Pour commencer, envisageons chaque catégorie l'une après l'autre – détenteur des données, spécialiste des données et tournure d'esprit orientée big data.

La chaîne de valeur des données

La substance de base des larges volumes de données, ce sont les informations elles-mêmes. Il est donc logique d'examiner d'abord les détenteurs de données. Ce ne sont pas forcément eux qui ont collecté les données au départ, mais ils en contrôlent l'accès, les utilisent eux-mêmes ou octroient une licence d'exploitation à des tiers qui en extraient la valeur. Par exemple, ITA Software, un grand réseau de réservation de billets d'avion (classé après Amadeus, Travelport et Sabre), a fourni des données à Farecast pour que cette société fasse ses prévisions de tarifs aériens, mais n'a pas effectué l'analyse lui-même. Pourquoi ? ITA, vu son activité, a considéré qu'il devait utiliser ses données pour son cœur de métier – la vente de billets d'avion – et non s'en servir pour des utilisations connexes. Ce qui l'aurait d'ailleurs obligé à contourner le brevet d'Etzioni.

Le réseau de réservations a également choisi de ne pas exploiter les données pour une autre raison : sa position dans la chaîne de valeur des informations. « Les gens d'ITA ont évité de se lancer dans des projets qui impliquaient un usage commercial des données trop étroitement lié aux recettes des compagnies aériennes, rappelle Carl de Marcken, cofondateur

et ancien directeur de la technologie d'ITA Software. ITA avait un accès privilégié à ces données, nécessaires aux services qu'elle proposait, et ne pouvait se permettre de compromettre cet atout. » Elle s'est donc tenue à distance en accordant une licence d'exploitation des données au lieu de les utiliser elle-même directement. La majeure partie de la valeur secondaire des données est revenue à Farecast – à ses clients sous forme de billets d'avion meilleur marché ainsi qu'à ses employés et propriétaires –, qui a tiré ses gains de publicités, de commissions et finalement de la vente de la société.

Certaines sociétés se sont astucieusement positionnées au centre des flux d'informations, de manière à acquérir de l'envergure et extraire de la valeur des données. Cela a été le cas du secteur des cartes de crédit aux États-Unis. Durant des années, le coût élevé de la lutte contre la fraude a conduit de nombreuses petites et moyennes entreprises bancaires à éviter d'émettre leurs propres cartes de crédit et à confier les activités qui y étaient liées à de grandes institutions financières, capables d'investir dans la technologie nécessaire vu leurs moyens. Des entreprises comme Capital One et MBNA de la Bank of America se sont emparées de ce marché. Mais les petites banques regrettent d'avoir pris cette décision, parce que la cession de ces activités les prive aujourd'hui de données qui les renseignent sur les habitudes de consommation de leurs clients et les empêche de leur vendre des services personnalisés.

Les grandes banques et les sociétés émettrices de cartes de crédit comme Visa et MasterCard semblent, elles, positionnées idéalement dans la chaîne de valeur des informations. Vu le service fourni à un grand nombre de banques et de commerçants, de nombreuses transactions circulent sur leurs réseaux et elles peuvent les utiliser pour en déduire le comportement des consommateurs. Leur modèle d'entreprise s'est converti du simple traitement des paiements à la collecte de données. La question est alors de savoir ce qu'elles en font.

MasterCard pourrait accorder une licence d'exploitation des données à des tiers qui en extrairaient la valeur, comme l'a fait ITA, mais la société préfère les analyser elle-même. Un service appelé MasterCard Advisors agrège et analyse 65 milliards de transactions effectuées par 1,5 milliard de titulaires de cartes dans 210 pays pour détecter les tendances du marché et de la consommation. Puis ces informations sont vendues à des tiers. La société a notamment découvert que si quelqu'un fait le plein vers seize

heures, il y a de fortes probabilités pour qu'il dépense entre 35 et 50 dollars dans l'heure qui suit à l'épicerie ou dans un restaurant. Un marketeur pourrait se servir de cette information pour imprimer au verso des reçus de station-service des bons pour un supermarché voisin qui seraient valables à ce moment de la journée.

Son statut d'intermédiaire dans les flux d'informations confère à MasterCard une position de premier plan pour collecter des données et en extraire de la valeur. On peut imaginer un avenir où les sociétés de cartes de crédit décideraient de renoncer à leurs commissions sur les transactions et de les traiter gratuitement en contrepartie d'un accès à une plus grande quantité de données. Elles tireraient alors leurs revenus de la vente d'analyses hautement sophistiquées de ces données.

La seconde catégorie se compose des spécialistes des données : des sociétés dotées de l'expertise ou des technologies nécessaires pour réaliser des analyses complexes. MasterCard a choisi de le faire en interne, et certaines entreprises oscillent entre les deux catégories. Mais beaucoup d'autres s'adressent à des spécialistes. Par exemple, le cabinet conseil Accenture travaille avec des entreprises de nombreux secteurs pour déployer des technologies avancées en matière de capteurs sans fil et analyser les données qu'ils collectent. Dans le cadre d'un projet pilote pour la ville de St. Louis, dans le Missouri, Accenture a installé des capteurs sans fil dans un certain nombre d'autobus publics pour effectuer la surveillance des moteurs, prédire les pannes ou déterminer le moment optimal pour effectuer un entretien. Résultat : une réduction des coûts de 10 %. À elle seule, une découverte a permis d'économiser plus d'un millier de dollars par véhicule : qu'il suffisait de changer une pièce tous les 450 000 kilomètres et non tous les 322 000 à 400 000 kilomètres comme cela était auparavant programmé. Et c'est le client, et non le cabinet conseil, qui a recueilli les fruits de ces données.

Autre exemple frappant, dans le monde des données médicales, où l'on voit la façon dont les entreprises de technologie peuvent fournir de précieux services. Le Med-Star Washington Hospital Center à Washington, D.C., en collaboration avec Microsoft Research, a utilisé son logiciel Amalga pour analyser des dossiers médicaux anonymisés, archivés sur plusieurs années – données démographiques des patients, analyses, diagnostics, traitements et autres. Et ce, dans le but de trouver comment réduire infections et taux de

réadmission, deux facteurs synonymes de frais d'hôpitaux majeurs. Tout bon résultat signifiant donc d'énormes économies.

La technique a révélé des corrélations surprenantes. Elle a fourni une liste de toutes les pathologies qui font augmenter le risque de voir un patient sorti de l'hôpital y revenir dans le mois qui suit. Certaines sont bien connues et la solution n'est pas toute trouvée. Un patient atteint d'une insuffisance cardiaque congestive a de fortes chances d'être hospitalisé de nouveau : c'est une affection difficile à soigner. Mais le système a également repéré une autre variable prédictive majeure inattendue : l'état psychologique du patient. La probabilité de réadmission d'une personne dans le mois suivant sa sortie est nettement plus forte si une détresse psychologique s'est traduite dans ses propos, lors de sa première hospitalisation, par l'utilisation de termes comme « dépression ».

Sans rien indiquer qui permette d'établir une relation de cause à effet, cette corrélation suggère néanmoins que suivre l'état de santé mentale du patient après sa sortie pourrait contribuer à améliorer sa santé physique, ce qui réduirait le nombre de réadmissions et diminuerait les frais médicaux. Cette découverte puisée par une machine dans une vaste mine de données, un être humain ne l'aurait sans doute jamais faite. La société Microsoft n'a pas contrôlé les données, qui appartenaient à l'hôpital ; le résultat n'a pas été le fruit d'une idée géniale (d'ailleurs pas nécessaire dans ce cas de figure) mais celui du logiciel Amalga, l'outil qu'elle a fourni pour détecter l'information qui a fini par en ressortir.

Les entreprises détentrices d'un volume important de données tablent sur des spécialistes pour en extraire de la valeur. Mais en dépit de leurs titres ronflants et des éloges dont ils font l'objet, la vie de ces « ninja des données », comme on appelle parfois ces experts techniques, n'est pas toujours aussi prestigieuse qu'il y paraît. Ils travaillent dur dans ces nouvelles mines de diamants que sont les big data, empochent de coquettes sommes, mais ils doivent remettre aux détenteurs des données les pierres précieuses qu'ils en ont extraites.

Dans le troisième groupe, on trouve des sociétés et des individus dont l'état d'esprit est orienté big data. Leur point fort est qu'ils voient les opportunités bien avant les autres – même s'ils ne disposent pas de données ni ne possèdent les compétences pour concrétiser ces opportunités. D'ailleurs, c'est sans doute précisément parce que leur esprit n'est pas

emprisonné derrière des barreaux imaginaires que ces « outsiders » ont une vision des possibilités au lieu d'être bloqués par l'idée qu'une chose n'est pas réalisable.

Bradford Cross est l'incarnation de l'état d'esprit orienté big data. En août 2009, alors âgé d'une vingtaine d'années, il crée avec quelques amis FlightCaster.com. Tout comme FlyOnTime.us, FlightCaster doit prévoir quel est le risque de retard d'un vol sur les lignes intérieures aux États-Unis. Pour établir ses prévisions, la société analyse alors tous les vols au cours des dix années précédentes, en les comparant aux données météorologiques historiques et actuelles.

Fait intéressant, les détenteurs de données eux-mêmes en étaient incapables. Aucun n'avait la motivation – ni le mandat réglementaire – pour utiliser les données de cette façon. En fait, si les sources de données – l'U.S. Bureau of Transportation Statistics^a, la Federal Aviation Administration^b et le National Weather Service^c – avaient osé prédire les retards sur des vols aériens commerciaux, le Congrès aurait probablement tenu audience et fait tomber des têtes de bureaucrates. Les compagnies aériennes ne pouvaient pas le faire – ou ne voulaient pas le faire. Elles ont en effet intérêt à rester le plus discrètes possible sur leurs performances quand celles-ci sont médiocres. Pour obtenir les prévisions de retard, il a fallu un groupe d'ingénieurs drapés du voile de l'anonymat. Les prévisions de FlightCaster se sont révélées si étonnamment exactes que les employés de la compagnie eux-mêmes se sont mis à les utiliser : les compagnies d'aviation sont réticentes à annoncer les retards avant la toute dernière minute, si bien que même si elles sont la source ultime d'informations, elles ne sont pas les plus à jour.

C'est grâce à son état d'esprit et à son idée géniale d'offrir au public une réponse attendue par des millions de gens, via un traitement de données accessibles, que Cross a fait de sa société FlightCaster la première sur ce créneau, quoique de justesse. Le mois du lancement de FlightCaster, les « geeks » qui dirigent FlyOnTime.us ont commencé à rassembler des données ouvertes pour réaliser leur site Web. L'avantage dont jouissait FlightCaster n'a pas tardé à se réduire. En janvier 2011, Cross et ses partenaires ont vendu la société à Next Jump, qui gère une gamme de programmes s'adressant aux entreprises et aux particuliers en utilisant des techniques de big data.

Puis Cross s'est tourné vers un autre secteur vieillissant et y a repéré un créneau dont un innovateur extérieur pouvait s'emparer : les médias d'information. Sa start-up, Prismatic, agrège et classe du contenu glané sur l'ensemble du Web en se fondant sur l'analyse de textes, les préférences des utilisateurs, la popularité liée aux réseaux sociaux et l'analyse de données massives. Fait important, le système ne fait pas une grande distinction entre un blog d'adolescent, un site Web d'entreprise et un article du *Washington Post* : si le contenu est jugé pertinent et s'il a du succès (mesuré au nombre de visites et de partages), il apparaît en haut de l'écran.

À travers le service qu'elle offre, Prismatic confère une reconnaissance au mode d'interaction des jeunes avec les médias. Pour cette génération, la source d'information a perdu son importance essentielle. Cela devrait donner une leçon d'humilité aux grands prêtres des grands médias en leur rappelant que, globalement, le public est encore mieux informé qu'eux, et que des journalistes en costume-cravate se retrouvent confrontés à des bloggeurs en peignoir. Le point essentiel est celui-ci : on aurait beaucoup de mal à imaginer qu'une société comme Prismatic émerge du secteur des médias lui-même, même si elle collecte énormément d'informations. Les habitués du bar du National Press Club n'ont jamais pensé à réutiliser des données en ligne au sujet de la consommation médiatique. Pas plus que des spécialistes des analyses d'Armonk, dans l'État de New York, ou de Bangalore, en Inde, n'auraient exploité les informations de cette manière. Et c'est Cross, un outsider au look un peu louche avec cheveux en broussaille et accent traînant de flemmard, qui a supposé qu'en utilisant des données il pouvait dire au public à quels sujets accorder de l'attention, mieux que les rédacteurs du *New York Times*.

La notion d'état d'esprit orienté big data et le rôle d'un outsider créatif et de sa brillante idée ne sont pas sans rappeler ce qui s'est produit au milieu des années 1990, à l'aube du commerce électronique : ses pionniers ne s'étaient alors pas encombres des idées arrêtées ou des garde-fous institutionnels des secteurs plus anciens. Ainsi, c'est un informaticien travaillant pour le compte d'un fonds d'investissement, Jeff Bezos, et non l'éditeur Barnes & Noble, qui a créé une librairie en ligne (Amazon). C'est un développeur de logiciels, Pierre Omidyar, et non le groupe de sociétés de ventes aux enchères Sotheby's, qui a créé un site de vente aux enchères (eBay). Aujourd'hui, les entrepreneurs avec un état d'esprit orienté big data ne disposent souvent pas de données au démarrage. Mais c'est justement

pour cette raison qu'ils sont dénués d'intérêts particuliers et ne sont pas soumis à des facteurs dissuasifs susceptibles de les empêcher de laisser libre cours à leurs idées.

Comme nous l'avons vu, il y a des cas où une entreprise combine un certain nombre de ces caractéristiques associées aux big data. Etzioni et Cross ont sans doute été des précurseurs avec leurs très bonnes idées, mais ils possédaient également les compétences. Les ouvriers de Teradata et Accenture ne se contentent pas de pointer ; ils sont connus pour avoir eux aussi une idée géniale de temps à autre. Néanmoins, les archétypes ont une utilité pour mesurer les rôles joués par différentes entreprises. Les pionniers actuels des big data viennent souvent d'horizons très divers et appliquent leurs compétences en matière de données dans un large éventail de domaines. Une nouvelle génération d'investisseurs providentiels et d'entrepreneurs émerge, notamment certains anciens de Google et de ce qu'on appelle la PayPal Mafia (anciens dirigeants de la société comme Peter Thiel, Reid Hoffman et Max Levchin). Ils sont, avec une poignée d'informaticiens universitaires, parmi les plus grands promoteurs de start-up actuelles dont l'activité repose totalement sur les données.

À l'aune de la créativité d'individus et d'entreprises œuvrant dans cette « chaîne alimentaire » des big data, nous pouvons réexaminer la valeur des sociétés. Par exemple, Salesforce.com n'est pas simplement une plateforme utile à des entreprises parce qu'elle héberge leurs applications : elle est également en bonne position pour exploiter la valeur des données qui circulent dans son infrastructure. Les entreprises de téléphonie mobile, comme nous l'avons vu dans le précédent chapitre, collectent une quantité gigantesque de données mais elles en ignorent souvent la valeur, de par leur culture d'entreprise. Il leur est toutefois possible d'octroyer une licence d'exploitation à des tiers capables d'en extraire une valeur nouvelle (tout comme Twitter, précédemment citée).

Certaines entreprises privilégiées sont à cheval sur les différents domaines par choix stratégique. Nous avons vu que Google collecte des données comme les fautes de frappe dans les requêtes et qu'elle a eu la brillante idée de les utiliser pour créer un correcteur orthographique avec l'avantage qu'elle dispose de compétences en interne pour concrétiser brillamment cette idée. Pour beaucoup de ses autres activités également, Google retire un avantage de son intégration verticale dans la chaîne de

valeur des big data, où elle occupe les trois positions à la fois. En parallèle, Google met aussi certaines de ses données à la disposition de tiers, via des interfaces de programmes d'application (API), de sorte que leur réutilisation apporte une valeur ajoutée. Les cartes de Google en sont un bon exemple ; elles sont utilisées sur l'ensemble du Web par tous, des agences immobilières aux sites Web gouvernementaux et ce, gratuitement (sauf pour les sites Web qui reçoivent beaucoup de visites).

Amazon, elle aussi, possède l'état d'esprit, l'expertise et les données. En fait, la société a abordé son modèle d'entreprise dans cet ordre, qui est à l'inverse de la norme. Son système de recommandation très reconnu maintenant n'était qu'une idée au départ. Son prospectus d'introduction en Bourse en 1997 décrivait le filtrage collaboratif » avant même « qu'Amazon ne sache comment le système fonctionnerait en pratique ou n'ait suffisamment de données pour qu'il soit utile.

Google et Amazon ont en commun de chevaucher les catégories, mais leurs stratégies diffèrent. Quand Google a entrepris de collecter toutes sortes de données, elle pensait déjà aux utilisations secondaires. Les voitures de son programme Street View, comme nous l'avons déjà vu, collectaient des informations, pas uniquement pour son service de cartographie mais aussi pour entraîner les voitures automatisées. En revanche, Amazon est plus axée sur l'utilisation initiale des données et les utilisations secondaires ne constituent qu'une source de revenu annexe. Son système de recommandation, par exemple, se sert des données du parcours de navigation comme d'un indicateur, mais la société n'a pas utilisé les informations pour réaliser des opérations extraordinaires comme faire des prédictions sur la situation économique ou sur des épidémies de grippe.

La liseuse de livres numériques Kindle commercialisée par Amazon est capable d'indiquer si le texte d'une page particulière a été beaucoup annoté et souligné par les utilisateurs, mais la société ne vend pas ces informations aux auteurs ni aux éditeurs. Les marketeurs aimeraient beaucoup savoir quels sont les passages très appréciés du plus grand nombre et utiliser ces informations pour mieux vendre les livres. Les auteurs aimeraient sans doute savoir à quel endroit de leurs imposants ouvrages la plupart des lecteurs en abandonnent la lecture, et pourraient se servir de l'information pour améliorer leur travail. Les éditeurs pourraient repérer les thèmes annonciateurs du prochain livre à succès. Mais Amazon semble laisser le terrain des données en jachère.

Habilement exploités, les gros volumes de données peuvent transformer les modèles d'entreprise des sociétés ainsi que les modes d'interaction entre partenaires de longue date. Un exemple stupéfiant est celui d'un grand constructeur automobile européen qui a redéfini ses relations commerciales avec un fournisseur de pièces détachées en exploitant des données concernant l'utilisation des véhicules dont l'équipementier ne disposait pas. (Cet exemple nous a été communiqué à titre purement informatif par l'une des principales entreprises qui a analysé les données, mais nous ne sommes malheureusement pas autorisés à dévoiler les noms des deux sociétés.)

Les véhicules actuels sont remplis de puces, de capteurs et de logiciels qui téléchargent, lors des visites d'entretien, les données relatives aux performances du véhicule dans les ordinateurs des constructeurs automobiles. Les voitures standard classiques sont maintenant dotés de quelque 40 microprocesseurs ; l'ensemble de l'électronique embarquée représente un tiers de leur coût, ce qui en fait les dignes successeurs des navires que Maury appelait « observatoires flottants ». La capacité de recueillir des données sur l'état d'usure effectif des pièces sur la route – et de réintégrer ces données pour les améliorer – est en passe de devenir un gros avantage compétitif pour les entreprises qui peuvent se procurer les informations.

Travaillant avec une société d'analyse externe, un constructeur automobile a pu détecter les dysfonctionnements d'un capteur placé dans le réservoir, fabriqué par un fournisseur allemand : il émettait plusieurs fausses alarmes pour une vraie. La société aurait pu transmettre cette information au fournisseur et exiger que le défaut du capteur soit rectifié. Du moins, c'est ce qui aurait pu se passer à une époque où régnait un peu plus de courtoisie dans les affaires. Mais le constructeur avait dépensé une fortune dans son programme d'analyse et il voulait se servir de cette information pour compenser une partie de son investissement.

Le constructeur réfléchit aux options possibles. Fallait-il vendre les données ? Comment les informations seraient-elles évaluées ? Et s'il se heurtait à des réticences de la part du fournisseur, et qu'il se retrouvait avec une pièce détachée toujours déficiente ? Il savait de surcroît que s'il transmettait l'information, les pièces similaires fournies à ses concurrents seraient améliorées elles aussi. Non, il fallait une manœuvre plus habile qui ne profite qu'à l'amélioration de ses propres véhicules. L'idée originale consista finalement à améliorer la pièce à l'aide d'un logiciel modifié. Une

technique qui a été brevetée puis vendue au fournisseur – ce qui a permis d’empocher une coquette somme au passage.

Les nouveaux intermédiaires en données

Quels sont les acteurs qui détiennent le plus de valeur dans la chaîne de valeur des big data ? Aujourd’hui, il semblerait que ce soit ceux qui possèdent l’état d’esprit, les idées innovantes. Comme nous l’avons vu à l’ère de la bulle Internet, les précurseurs peuvent vraiment connaître la prospérité. Mais leur avantage ne dure pas forcément très longtemps. Plus nous avancerons dans l’ère des big data, plus nombreux seront ceux qui adopteront son état d’esprit, et l’avantage des pionniers du domaine diminuera comparativement.

Alors, l’essentiel de la valeur résiderait-il dans les compétences ? Après tout, une mine d’or ne vaut rien si on n’en extrait pas l’or. Et pourtant, l’histoire de l’informatique laisse penser le contraire. Aujourd’hui, l’expertise en matière de gestion de données, de science des données, d’analyse, d’algorithmes d’apprentissage machine et autres sont en forte demande. Mais avec le temps, quand les big data feront de plus en plus partie intégrante de notre vie quotidienne, que les outils se seront améliorés et seront plus faciles à utiliser, que plus nombreux seront ceux ayant acquis de l’expertise dans ce domaine, la valeur des compétences diminuera aussi en termes relatifs. De la même manière, la compétence en matière de programmation informatique s’est généralisée entre les années 1960 et 1980. Aujourd’hui, les entreprises de sous-traitance offshore ont fait perdre encore plus de valeur à la programmation ; ce qui était auparavant le modèle du savoir-faire technique est maintenant un moteur de développement des populations pauvres du monde. Cela ne veut pas dire que l’expertise en matière de big data revêt peu d’importance. Mais ce n’est pas la source de valeur la plus essentielle, puisqu’il est possible de l’importer de l’extérieur.

Aujourd’hui, dans ces premiers stades des big data, ce sont les idées et les compétences qui semblent représenter la valeur la plus importante. Sachant qu’à terme, elle finira par se retrouver en majeure partie dans les données elles-mêmes – parce que d’une part, on saura mieux exploiter les informations, et d’autre part, les détenteurs de données sauront mieux jauger la valeur potentielle de l’atout en leur possession. Par conséquent, ils

y tiendront comme à la prune de leurs yeux, et factureront des prix élevés pour en donner l'accès à des acteurs extérieurs. Enfin, pour filer la métaphore de la mine d'or, nous dirons que c'est l'or lui-même qui demeurera l'élément essentiel.

Une dimension importante mérite toutefois d'être relevée : celle de l'émergence de détenteurs de données sur le long terme. Dans certains cas, apparaîtront des « intermédiaires en données » capables de collecter les données de multiples sources, de les agréger et d'en extraire des produits novateurs. Les détenteurs de données laisseront jouer ce rôle à des intermédiaires parce que ce sera le seul moyen d'exploiter une certaine partie de la valeur des données.

Prenons comme exemple Inrix, une société d'analyse de trafic au large de Seattle^d. Elle compile des données de géolocalisation en temps réel provenant de 100 millions de véhicules en Amérique du Nord et en Europe, des BMW, Ford et Toyota, entre autres, ainsi que de véhicules commerciaux comme les taxis et les camionnettes de livraison. Elle obtient également les données issues de téléphones portables d'automobilistes individuels (ses applications smartphone gratuites jouent ici un rôle important : les utilisateurs obtiennent des informations routières et Inrix récupère en contrepartie leurs coordonnées). Inrix combine ces informations avec des données sur l'historique des conditions de circulation, le temps et d'autres éléments comme les événements locaux pour donner des prévisions sur la circulation. Le produit de sa ligne d'assemblage de données est transmis au système de navigation des véhicules et les parcs de véhicules commerciaux ainsi que ceux du gouvernement l'utilisent.

Inrix représente la quintessence de l'intermédiaire en données indépendant. Elle collecte ses informations auprès de nombreux constructeurs automobiles concurrents et génère ainsi un produit dont la valeur est supérieure à tout ce qu'aurait pu obtenir n'importe lequel d'entre eux à lui tout seul. Les quelques millions de points de données provenant de ses véhicules en circulation que détient chaque constructeur automobile lui permettraient de prédire les flux de circulation routière, mais ces prédictions ne seraient ni très précises, ni très complètes. La qualité des prédictions s'améliore avec l'augmentation de la quantité de données. De surcroît, les constructeurs automobiles n'ont sans doute pas les compétences adéquates : les leurs consistent surtout à plier du métal et non à cogiter sur la loi de Poisson^e. Ils ont donc tous une bonne raison de s'adresser à un tiers pour

effectuer cette tâche. Par ailleurs, si les prévisions sur la circulation importent aux automobilistes, elles influencent rarement la décision d'acheter tel ou tel modèle de voiture. Aussi les concurrents ne voient-ils pas d'obstacle à unir ici leurs forces.

Naturellement, dans de nombreux secteurs, les sociétés partagent depuis longtemps des informations, notamment les laboratoires au service des assureurs et les secteurs travaillant en réseau comme la banque, l'énergie et les télécoms. L'échange d'informations y est primordial pour éviter les problèmes et, parfois, il est exigé par les organismes de réglementation. Les sociétés d'étude de marché agrègent des données depuis des décennies, comme le font les sociétés créées pour assurer des tâches spécialisées tel le contrôle de la diffusion de journaux. Pour certaines associations professionnelles, il s'agit de leur cœur de métier.

La différence est qu'aujourd'hui les données sont devenues une matière première qui arrive sur le marché ; une ressource indépendante de ce qu'elles étaient destinées à mesurer auparavant. Par exemple, les informations d'Inrix sont plus utiles qu'il n'y paraît à première vue. Son analyse de la circulation est utilisée pour mesurer la santé des économies locales parce qu'elle donne des indications sur l'emploi, les ventes au détail et les loisirs. Quand la reprise économique aux États-Unis a eu des ratés en 2011, c'est par l'analyse de la circulation qu'on a pu le déceler, en dépit des démentis des responsables politiques à ce sujet : il y avait moins d'embouteillages aux heures de pointe, ce qui indiquait une hausse du taux de chômage. Par ailleurs, Inrix a vendu ses données à un fonds d'investissement qui utilise celles concernant les conditions de circulation aux alentours des magasins d'une grande chaîne de distribution comme indicateur des ventes ; le fonds s'en sert pour négocier les actions de la société avant son annonce de revenus trimestriels. Plus de voitures égale de meilleures ventes.

D'autres intermédiaires de ce type apparaissent dans la chaîne de valeur des big data. Un des premiers a été Hitwise, racheté plus tard par Experian, qui a passé des accords avec des fournisseurs d'accès à Internet pour collecter leurs données de navigation en contrepartie de revenus supplémentaires. Les données ont été concédées sous licence à un petit prix forfaitaire plutôt que contre un pourcentage de la valeur générée par ces données. Hitwise a récupéré la plus grande partie de la valeur en tant qu'intermédiaire. Un autre exemple est Quantcast, qui mesure le trafic en

ligne sur des sites Web de sociétés pour aider ces dernières à en savoir plus sur les données démographiques et le mode d'utilisation de leurs visiteurs. Elle met gracieusement à la disposition des sites un outil en ligne pour suivre les visites ; en contrepartie, Quantcast a la possibilité de voir les données, ce qui lui permet d'optimiser son ciblage publicitaire.

Ces nouveaux intermédiaires ont identifié des créneaux de marché lucratifs qui ne présentent pas de menace pour le modèle d'entreprise des détenteurs de données auprès de qui ils obtiennent leurs données. Pour le moment, la publicité sur Internet est l'un de ces créneaux, du fait que c'est là que se trouve la plupart des données et qu'il existe un besoin impérieux de les exploiter pour cibler les annonces publicitaires. Mais avec la généralisation de la mise en données à de plus en plus de secteurs et la prise de conscience croissante que leur cœur de métier consiste à extraire des informations de ces données, ces intermédiaires indépendants feront leur apparition dans d'autres secteurs également.

Tous ces intermédiaires ne sont pas forcément des entreprises commerciales, ce peuvent être aussi des organisations à but non lucratif. Par exemple, le Health Care Cost Institute a été créé en 2012 par une poignée de compagnies d'assurance maladie parmi les plus grandes des États-Unis. La totalité de toutes leurs données cumulées se monte à cinq milliards de demandes d'indemnisation (anonymisées) provenant de 33 millions de personnes. C'est grâce au partage des informations que les sociétés peuvent repérer les tendances qui ne seraient sans doute pas ressorties de leurs ensembles de données plus petits. Une des premières conclusions a été que les frais médicaux aux États-Unis ont augmenté trois fois plus vite que l'inflation en 2009-2010, mais avec des différences marquées si l'on y regarde de près : les tarifs des urgences ont augmenté de 11 % alors que ceux des soins infirmiers ont en fait baissé. De toute évidence, les compagnies d'assurances n'auraient jamais confié leurs précieuses données à qui que ce soit d'autre qu'un intermédiaire dont le statut était à but non lucratif. Il y a moins de suspicion sur les intentions de ce type d'organisation, et à sa création peuvent être intégrées des valeurs de transparence et de responsabilité.

La diversité des sociétés dont l'activité concerne les big data montre bien l'évolution de la valeur des informations. Dans le cas de Decide.com, les données sur les prix sont fournies par des sites Web partenaires sur la

base d'un partage des recettes. Decide.com encaisse des commissions sur les produits achetés par les visiteurs via le site, mais les sociétés qui ont fourni les données obtiennent elles aussi une part du gâteau. C'est le signe d'une évolution dans la manière dont le secteur exploite commercialement les données : par le passé, l'ITA ne recevait pas de commissions sur les données qu'elle fournissait à Farecast, mais uniquement une redevance de licence de base. Maintenant, les fournisseurs de données arrivent à convenir de conditions plus intéressantes. Pour sa prochaine start-up, on peut supposer qu'Etzioni essaiera de fournir les données lui-même, du fait que la valeur se déplace maintenant vers les données après être passée de l'expertise à l'idée.

Le déplacement de la valeur vers ceux qui contrôlent les données entraîne des bouleversements dans les modèles d'entreprise. Le constructeur automobile qui a passé l'accord avec son fournisseur sur la propriété intellectuelle disposait en interne d'une solide équipe d'analyse de données mais avait besoin de collaborer avec un fournisseur de technologie extérieur pour comprendre ce que les données lui apportaient comme éclairage. La société de technologie a été rémunérée pour son travail, mais le constructeur automobile a conservé pour lui l'essentiel des profits. Flairant l'opportunité, la société de technologie a alors légèrement modifié son modèle d'entreprise pour partager les risques et les avantages avec les clients. Elle est convenue d'honoraires plus faibles avec en contrepartie la récupération d'une part de la valeur produite par son analyse. (Quant aux fournisseurs de pièces détachées, il est fort probable qu'à l'avenir, ils voudront tous ajouter des capteurs de mesure à leurs produits, ou insisteront pour intégrer dans le contrat de vente une clause qui leur donne accès aux performances, afin de pouvoir améliorer leurs composants en permanence.)

Quant aux intermédiaires, leur vie est compliquée parce qu'ils doivent convaincre les sociétés de l'intérêt du partage. Par exemple, Inrix a commencé à collecter d'autres informations que celles concernant la géolocalisation. En 2012, il a analysé à titre expérimental où et quand les systèmes automatiques de freinage (ABS) subissaient des dommages ; cela pour le compte d'un constructeur automobile qui a conçu un système de télémétrie pour la collecte d'informations en temps réel. L'hypothèse à la base de cette analyse est que le déclenchement fréquent de l'ABS sur un tronçon de route particulier peut indiquer que l'état de la route est dangereux à cet endroit et que les automobilistes devraient envisager

d'autres itinéraires. Ainsi, à l'aide de ces données, Inrix s'est retrouvé en mesure de recommander non seulement le plus court itinéraire mais aussi le plus sûr.

Mais le constructeur automobile n'a pas voulu partager ces données avec d'autres. Il a exigé d'Inrix qu'il ne déploie le système que dans ses véhicules. Pouvoir se vanter de disposer d'un tel dispositif apporterait une valeur qui l'a visiblement emporté pour le constructeur sur le profit à tirer de l'agrégation de ses données avec celles de tiers en vue d'améliorer la précision globale du système. Cela dit, Inrix reste convaincu qu'à terme, tous les constructeurs automobiles verront l'utilité d'agréger toutes leurs données. En tant qu'intermédiaire en données, Inrix a de bonnes raisons de s'entêter dans son optimisme : son activité repose entièrement sur l'accès à de multiples sources de données.

Pour exercer une activité dans le domaine des big data, les entreprises font par ailleurs l'expérience de différentes structures organisationnelles. Le modèle d'entreprise choisi par Inrix n'est pas le fruit du hasard, comme c'est le cas lors de la création d'un bon nombre de start-up – son rôle d'intermédiaire est un choix délibéré. Microsoft, qui détenait les brevets essentiels liés à cette technologie, a pensé qu'une petite entreprise indépendante – plutôt qu'une grosse société – serait perçue comme plus neutre, pourrait rassembler des concurrents du secteur et tirer le maximum de sa propriété intellectuelle. De même, le MedStar Washington Hospital Center qui a utilisé le logiciel de Microsoft Amalga pour analyser les réadmissions de patients savait pertinemment ce qu'il faisait de ses données : le système Amalga était à l'origine le logiciel de la salle d'urgence appartenant à l'hôpital lui-même, appelé Azyxxi, qu'il avait vendu en 2006 à Microsoft pour qu'elle le perfectionne.

En 2010, UPS a vendu au fonds d'investissement privé Thoma Bravo une unité d'analyse de données en interne, appelée UPS LogisticsTechnologies. L'unité qui opère maintenant sous le nom de Roadnet Technologies est libre de répondre à la demande d'autres sociétés et d'effectuer des analyses d'itinéraires pour elles. La collecte de données auprès de nombreux clients lui permet de fournir un service d'analyse comparative portant sur l'ensemble du secteur, service d'ailleurs utilisé tant par UPS que par ses concurrents. Du temps où elle était UPS Logistics, elle n'aurait jamais réussi à persuader les concurrents de sa société mère de lui confier leurs ensembles de données, explique le directeur général de

Roadnet, Len Kennedy. Mais une fois qu'elle a été indépendante, les concurrents d'UPS se sont sentis plus à l'aise pour lui fournir leurs données et, finalement, l'amélioration de la précision apportée par l'agrégation des données a pu bénéficier à tous.

La preuve que ce sont les données elles-mêmes, plutôt que les compétences ou l'état d'esprit, qui sont en passe de posséder la plus grande valeur est visible dans les nombreuses acquisitions d'entreprises spécialisées en big data. Par exemple, en 2006, Microsoft a récompensé l'état d'esprit orienté big data d'Etzioni en achetant Farecast pour près de 110 millions de dollars. Mais, deux ans plus tard, Google a payé 700 millions pour acquérir le fournisseur de données de Farecast, ITA Software.

La fin de l'expert

Dans le film *Le Stratège*, qui raconte comment l'Oakland Athletics est devenue une équipe de baseball gagnante grâce à l'application à ce sport de l'analyse et de nouveaux types de mesures, il y a une scène savoureuse dans laquelle de vieux dénicheurs de talents grisonnants, assis autour d'une table, discutent des joueurs. Le public ne peut s'empêcher de tiquer, pas simplement parce que la scène dévoile comment les décisions sont prises sans s'appuyer sur aucune donnée, mais parce que nous nous sommes tous retrouvés dans des situations où la « certitude » était fondée sur des impressions plutôt que sur la rationalité.

« Il a un corps fait pour le baseball... et il a une belle gueule, dit un des dénicheurs.

— Il a un bon swing. Quand la balle entre en contact avec la batte, son mouvement pour accompagner la balle est magnifique, la balle décolle carrément de la batte, renchérit un homme frêle, aux cheveux gris portant une prothèse auditive.

— Oui, carrément, approuve un autre.

Un troisième homme interrompt la conversation en déclarant : « Sa petite amie est moche.

— Ça prouve quoi ? demande le dénicheur de talents qui dirige la réunion.

— Une petite amie moche, c'est un signe de manque de confiance, explique tout bonnement le détracteur de service.

— D'accord », dit le chef du groupe, satisfait et prêt à poursuivre.

Après un échange de propos sur le ton de la plaisanterie, l'un d'eux, silencieux jusque-là, prend la parole : « Ce type a du caractère, et ça, c'est bon pour nous. Ce que je veux dire, c'est que ce gars, quand il entre dans la pièce, sa queue y était déjà à l'œuvre depuis deux minutes. » Un autre ajoute : « Il peut décrocher le premier prix dans un concours de beaux gosses. Il est prêt à faire le boulot. Il a juste besoin de jouer quelques matchs.

— Je dis simplement, reprend le détracteur, que je ne donnerais pas à cette nana plus qu'un six sur dix ! »

Cette scène dépeint parfaitement les failles du jugement humain. Ce qui passe pour un débat raisonné n'est en réalité fondé sur rien de concret. Alors que des millions de dollars sont en jeu pour des contrats de joueurs, les décisions se fondent sur l'instinct, en l'absence de toute mesure objective. Certes, ce n'est qu'un film, mais dans la vie réelle, ce n'est pas très différent. Ce même type de raisonnement creux est employé dans d'autres contextes, des salles de conseils d'administration de Manhattan au Bureau ovale de Washington et jusqu'aux comptoirs et cuisines partout ailleurs.

Le Stratège, fondé sur le livre de Michael Lewis, narre l'histoire véridique de Billy Beane. Manager des Oakland Athletics, il a jeté aux orties le règlement centenaire sur les méthodes d'évaluation des joueurs et a opté pour une méthode fondée sur les mathématiques afin d'examiner ce sport à l'aide d'un nouvel ensemble de paramètres. Exit les traditionnelles statistiques comme la « moyenne au bâton » et bonjour la façon apparemment bizarre de penser au sport en termes de « moyenne de présence sur les buts ». La méthode fondée sur les données a révélé une dimension de ce sport qui avait toujours existé mais qui demeurait enfouie dans les cacahuètes et pop corn des spectateurs. Peu importe la manière dont un joueur se rend sur les buts, que ce soit par une balle rebondissante ou un but-sur-balles honteux, du moment qu'il s'y est rendu. Quand les données indiquent que voler des buts est une action inefficace, exit l'un des éléments du jeu les plus excitants mais les moins « productifs ».

En dépit de l'ampleur de la controverse, Beane a mis en place au sein de la direction de l'équipe la méthode connue sous le nom de « sabermétrie », terme créé par l'historien du baseball Bill James en référence à la Society for American Baseball Research (SABR), qui était

jusque-là le domaine d'une sous-culture geek. Beane a révolutionné les pratiques de recrutement du baseball, tout comme Galilée a adopté la vision héliocentrique en se heurtant jadis à l'autorité de l'Église catholique. Finalement, il a mené l'équipe, restée très longtemps en difficulté, à la première place dans la Ligue américaine Ouest durant la saison 2002, avec une série de 20 victoires. Depuis lors, les statisticiens ont supplanté les dénicheurs de talents dans le rôle de savants du milieu du sport. Et bien d'autres équipes se sont ruées pour adopter elles aussi la sabermétrie.

Dans le même esprit, le plus grand impact des big data se mesurera quand le jugement humain sera complété, voire supplanté par des décisions fondées sur des données. Dans son ouvrage *Super Crunchers*, l'économiste et professeur de droit de Yale, Ian Ayres, soutient que les analyses statistiques obligent les gens à reconsidérer leur instinct. Et quand il s'agit d'analyses à partir d'énormes masses de données, cela devient encore plus indispensable. Les experts en la matière, les spécialistes vont perdre une partie de leur lustre, comparé au statisticien et à l'analyste de données qui sont, eux, affranchis des anciennes méthodes et qui laissent les données parler. Ce nouveau cadre s'appuiera sur les corrélations sans jugements préconçus ni préjugés, tout comme Maury ; celui-ci n'a pas pris pour argent comptant ce que racontaient de vieux capitaines autour d'une bière dans un pub évoquant tel ou tel passage maritime, il a fait confiance aux données agrégées pour révéler des vérités pratiques.

Nous observons le déclin de l'influence des experts dans de nombreux domaines. Dans les médias, le contenu créé et affiché sur des sites Web comme Huffington Post, Gawker et Forbes est régulièrement déterminé par les données, pas uniquement par le jugement des rédacteurs. Les données révèlent ce que le public veut lire quasiment mieux que l'intuition de journalistes chevronnés. La société d'enseignement en ligne Coursera utilise des données indiquant quelles parties d'un cours vidéo les étudiants ont réécoutées de façon à savoir quel contenu n'était pas clair, et pour répercuter les informations aux enseignants afin de les aider à améliorer leur cours. Comme nous l'avons noté plus haut, Jeff Bezos a licencié les salariés d'Amazon qui écrivaient les critiques de livres quand les données ont montré que les recommandations générées par des algorithmes augmentaient les ventes.

Autrement dit, les compétences nécessaires pour réussir sur le lieu de travail sont en mutation. Cela modifie les attentes des organisations vis-à-

vis des contributions de leurs employés. Le Dr McGregor, qui s'occupe de bébés prématurés dans l'Ontario, n'a pas besoin d'être le médecin le plus chevronné de l'hôpital, ni l'autorité la plus éminente du monde en matière de soins néonataux, pour obtenir les meilleurs résultats pour ses patients. En fait, elle n'est pas docteur en médecine du tout – elle possède un doctorat en informatique. Mais elle se sert de données correspondant à plus d'une décennie de patients-années, que l'ordinateur analyse et qu'elle convertit en prescriptions médicales.

Comme nous l'avons vu, les pionniers des big data viennent souvent de domaines autres que celui dans lequel ils se taillent une place. Ils sont spécialistes en analyse de données, en intelligence artificielle, en mathématiques ou en statistiques, et ils mettent leurs compétences au service de secteurs spécifiques. Les lauréats des concours organisés par la société Kaggle sur sa plateforme en ligne, où des sociétés viennent exposer leurs projets liés aux big data, sont généralement nouveaux dans le secteur où ils créent des algorithmes efficaces, explique Anthony Goldbloom, directeur général de Kaggle. Un physicien britannique a failli remporter le concours avec les algorithmes qu'il avait créés pour prédire des déclarations de sinistre et détecter les voitures d'occasion défectueuses. Un actuaire de Singapour a lancé un concours pour prédire les réactions biologiques à des composés chimiques. Dans le même temps, les ingénieurs du groupe de traduction automatique de Google fêtent les traductions qu'ils ont mises au point de langues que personne ne parle dans le service ! De même, les statisticiens de l'unité de traduction automatique de Microsoft aiment bien ressortir une vieille blague : à savoir que la qualité des traductions s'améliore toutes les fois qu'un linguiste quitte l'équipe !

Bien sûr, les experts ne sont pas près de disparaître. Mais leur suprématie va décliner. Dorénavant, ils devront partager la tribune avec les accros des big data, tout comme la « reine » causalité doit partager la vedette avec l'humble corrélation. Cela transforme notre façon d'évaluer la connaissance, parce que nous avons plutôt eu tendance à penser que les spécialistes très pointus valaient mieux que les généralistes – que la fortune souriait à ceux qui maîtrisaient leur domaine. Pourtant, l'expertise est semblable à l'exactitude : elle est adaptée à un monde de petits volumes de données où l'on manque d'informations ou bien de bonnes informations, et où l'on doit donc compter sur l'intuition et l'expérience pour se guider. Dans un tel contexte, l'expérience joue un rôle essentiel, puisqu'elle est le

résultat de connaissances latentes accumulées sur une longue période – connaissances difficiles à transmettre ou à acquérir de manière livresque, ce dont l'on n'est même pas conscient – qui permettent de prendre des décisions plus judicieuses.

Mais quand on dispose de données à gogo, il est possible d'exploiter cette mine, et avec de meilleurs résultats. Ceux qui ont sous la main des big data à analyser n'auront sans doute pas de mal à dépasser les croyances et idées reçues non pas parce qu'ils sont plus intelligents, mais parce qu'ils détiennent les données. (De plus, étant des « outsiders », ils conservent une neutralité face aux querelles intestines, qui rendent la vision des experts étriquée.) Cela indique que les qualités attendues d'un employé varient d'une société à l'autre – le type de connaissances à posséder, les personnes à connaître et les domaines à étudier pour se préparer à la vie professionnelle.

Mathématiques et statistiques, avec peut-être une pincée de programmation et de science des réseaux seront aussi fondamentales dans le monde moderne du travail que calculer au siècle dernier et savoir lire et écrire à des époques antérieures. Par le passé, pour être un excellent biologiste, il fallait connaître d'autres biologistes. Cela n'a pas complètement changé. Mais, aujourd'hui, une profonde expertise dans un domaine ne suffit plus, l'éventail offert par les big data compte également. Il est tout à fait possible de résoudre un casse-tête dans le domaine de la biologie en s'associant avec un astrophysicien ou un concepteur de visualisation de données.

Les jeux vidéo sont un domaine où les lieutenants des big data se sont déjà taillé une place aux côtés des généraux de l'expertise, et le secteur s'en est trouvé métamorphosé. Il compte maintenant de grandes entreprises et ses recettes annuelles mondiales sont supérieures à celles du box office d'Hollywood. Auparavant, les sociétés concevaient un jeu, le lançaient en espérant qu'il serait un succès. En fonction des données sur les ventes, les sociétés préparaient une suite ou un nouveau jeu. Les décisions concernant le rythme du jeu et ses composants – personnages, scénario, objets et événements – reposaient sur la créativité des concepteurs, lesquels mettent à la tâche le même sérieux que Michel-Ange peignant la chapelle Sixtine. C'était de l'art et non de la science ; un univers d'intuitions et d'instincts, un peu comme celui des découvreurs de talents du film *Le Stratège*.

Cette époque est révolue. Les jeux FarmVille, FrontierVille, FishVille, créés par la société Zynga, ainsi que d'autres jeux interactifs se trouvent en ligne. À première vue, les jeux en ligne permettent à Zynga de regarder comment les joueurs s'en servent et de les modifier en fonction de cela. Si ces derniers rencontrent des difficultés pour sauter d'un niveau à l'autre, ou ont tendance à abandonner à un certain moment parce que l'action ralentit, Zynga peut repérer ces problèmes dans les données et peut y remédier. Mais, chose moins évidente, la société peut adapter les jeux en fonction des traits de caractère personnels des joueurs. Il n'y a donc pas une unique version de FarmVille, mais des centaines !

Chez Zynga, les analystes de big data étudient les ventes des marchandises virtuelles pour déterminer si elles sont influencées par la couleur, ou par le fait que les joueurs voient leurs amis les utiliser. Par exemple, quand les données ont montré que les joueurs de FishVille avaient acheté un poisson translucide six fois plus souvent que d'autres animaux, Zynga a proposé un nombre plus grand d'espèces translucides et en a tiré profit. Dans le jeu MafiaWars, les données ont révélé que les joueurs achetaient plus d'armes après que la société y eut ajouté une bordure dorée et plus de tigres quand elle a introduit dans le jeu des tigres blancs.

Un concepteur de jeux travaillant en studio n'aurait jamais pu se rendre compte de ces détails, tandis que les données les ont pointés. « Nous sommes une société spécialisée dans l'analyse de données déguisée en éditeur de jeux vidéo. Tout est commandé par les chiffres », expliquait Ken Rudin, alors responsable de l'analyse des données chez Zynga, avant de quitter la société pour diriger le service analyse de données chez Facebook. Exploiter des données ne garantit pas la réussite d'une entreprise mais montre que cela est possible.

Le changement pour une prise de décisions fondée sur les données est profond. La plupart des personnes appuient leurs décisions sur une combinaison de faits et de réflexion, auxquels s'ajoute une bonne dose de conjectures. « Une profusion de visions subjectives – des sensations dans le plexus solaire », pour reprendre les mots mémorables du poète W. H. Auden. Thomas Davenport, professeur de commerce au Babson College dans le Massachusetts et auteur de nombreux ouvrages sur l'analyse, appelle cela la « précieuse intuition ». Les dirigeants d'entreprises sont absolument sûrs d'eux-mêmes par instinct, et ils le suivent. Mais cela commence à changer maintenant que les décisions managériales sont plus

souvent prises ou du moins confirmées par l'analyse d'énormes masses de données et une modélisation prédictive.

Par exemple, The-Numbers.com, en utilisant de gros volumes de données et pléthore de calculs, peut dire aux producteurs du cinéma indépendant d'Hollywood, bien avant le tournage de la première scène, quelles recettes un film est susceptible de faire. La base de données de la société analyse près de 30 millions d'informations couvrant tous les films commerciaux américains sur plusieurs décennies. Cela inclut pour chacun le budget, le genre, les acteurs, l'équipe de tournage et les récompenses obtenues, ainsi que les recettes (au box office américain et international, droits étrangers, ventes et locations de vidéos, etc.), et bien plus encore. La base de données inclut également un réseau de relations humaines, par exemple « ce scénariste a travaillé avec ce réalisateur ; ce réalisateur a travaillé avec cet acteur », explique son fondateur et président, Bruce Nash.

The-Numbers.com est capable de trouver des corrélations complexes qui prédisent les futures recettes de films au stade de projet. Les producteurs présentent ces informations aux studios et aux investisseurs pour obtenir un soutien financier. La société peut même remanier le choix de variables pour dire aux clients comment augmenter leurs gains (ou minimiser le risque de pertes). Dans un cas, son analyse a trouvé qu'un projet aurait bien plus de chances de rencontrer le succès si la vedette masculine faisait partie du classement des acteurs les plus cotés : plus précisément encore, un acteur déjà nommé pour les Oscars et dont les cachets peuvent atteindre 5 millions de dollars. Dans un autre cas, Nash a informé le studio IMAX qu'un documentaire sur la navigation à voile ne serait rentable que si son budget évalué à 12 millions de dollars était réduit à 8 millions de dollars. « Le producteur était content – mais le réalisateur l'était moins », a commenté Nash.

Qu'il s'agisse de décider ou pas de faire un film ou d'engager tel ou tel sportif, le changement opéré dans le mode de prise de décision des entreprises commence à avoir des répercussions visibles sur les résultats financiers. Erik Brynjolfsson, professeur de commerce à la Sloan School of Management du MIT, et ses collègues ont étudié la performance des sociétés qui excellent dans la prise de décisions fondée sur les données et l'ont comparée avec celle d'autres sociétés. Ils ont constaté que la productivité était jusqu'à 6 % supérieure dans les premières que dans les secondes où aucun accent n'était mis sur les données pour prendre des

décisions. Cela confère aux entreprises qui en tiennent compte un avantage conséquent – même si, à l’instar de l’avantage de l’état d’esprit et des compétences, il ne devrait pas durer, de plus en plus de sociétés adoptant cette approche big data.

Une question de service

En devenant une source d’avantage concurrentiel pour de nombreuses sociétés, les big data vont remodeler la structure de secteurs entiers. Mais les avantages ne seront pas aussi importants pour tous. Les vainqueurs se trouveront dans les grandes et dans les petites entreprises mais, entre les deux, la masse sera évincée.

Les acteurs les plus importants comme Amazon et Google continueront à prospérer. Mais à l’inverse de la situation qui prévalait à l’ère industrielle, leur avantage concurrentiel ne reposera pas sur la taille physique. Le contrôle de la vaste infrastructure technique des centres de données est important mais ce n’est pas leur atout le plus essentiel. Parce que de grands volumes de stockage et de traitement numériques peuvent être loués à bon marché et augmentés en quelques minutes, les sociétés peuvent adapter leur puissance de calcul et leur capacité de stockage en fonction de la demande réelle. L’avantage dont jouissaient les grandes sociétés en termes de taille fondée sur l’infrastructure technique a diminué quand un coût fixe s’est transformé en coût variable. La taille a encore de l’importance mais elle a changé de nature.

Ce qui compte maintenant, c’est la taille des bases de données ; autrement dit détenir de vastes ensembles de données et pouvoir en recueillir toujours plus sans difficulté. Ainsi, les détenteurs de vastes volumes de données vont prospérer grâce à la collecte et au stockage d’encore plus grandes quantités de ce qui constitue la matière première de leur activité et qu’ils pourront réutiliser pour créer davantage de valeur.

L’enjeu pour les vainqueurs dans un monde de petits volumes de données ainsi que pour les grandes entreprises hors ligne – des sociétés comme Walmart, Procter & Gamble, GE, Nestlé et Boeing – est d’évaluer le pouvoir des big data, de collecter et utiliser les données de la manière la plus judicieuse. Le fabricant de moteurs d’avion Rolls-Royce a complètement transformé son activité au cours de ces dix dernières années en faisant l’analyse des données issues de ses produits au lieu de

simplement les fabriquer. À partir de son centre des opérations en Grande-Bretagne, la société surveille en permanence la performance de plus de 3 700 réacteurs à travers le monde pour détecter des problèmes avant que les pannes ne surviennent. Elle s'est servie de données pour transformer son modèle d'entreprise de la fabrication simple à celui des produits liés : Rolls-Royce vend des moteurs mais propose également de les contrôler ; la facturation à ses clients est fondée sur le temps d'utilisation (elle répare et remplace les moteurs en cas de problème). Les services représentent maintenant près de 70 % du chiffre d'affaires annuel de la division aéronautique civile.

Les start-up tout autant que les vieux piliers dans de nouveaux secteurs d'activité se positionnent pour recueillir de vastes flux de données. L'incursion d'Apple dans le marché des téléphones mobiles en est un bon exemple. Avant l'arrivée de l'iPhone, les opérateurs de téléphonie mobile amassaient des données en provenance des abonnés mais n'en tiraient pas profit. Par contre, Apple a exigé dans ses contrats avec des opérateurs de recevoir la plus grande partie des informations les plus utiles. Grâce aux données récupérées auprès de nombreux opérateurs à travers le monde, Apple obtient un tableau bien plus complet de l'utilisation des téléphones portables que ne le ferait un opérateur de téléphonie mobile à lui tout seul.

Les big data offrent des possibilités formidables à l'autre extrémité de l'éventail des tailles d'entreprises. Des petites sociétés astucieuses et réactives peuvent bénéficier de la « taille sans la masse », selon la célèbre expression du professeur Brynjolfsson. Autrement dit, elles peuvent avoir une présence virtuelle conséquente sans disposer de ressources physiques importantes, et elles peuvent diffuser très largement des innovations à moindre coût. Fait important, comme certains des meilleurs services en matière de big data reposent essentiellement sur des idées novatrices, il n'y a pas forcément besoin d'un gros investissement au départ. Les petites entreprises peuvent utiliser des données sous licence plutôt qu'en être propriétaires, effectuer leurs analyses sur une plateforme d'informatique dématérialisée pour un faible coût et payer les droits de licence sur un pourcentage du revenu généré.

Il y a de fortes chances que ces avantages aux deux bouts de l'éventail ne se limiteront pas aux utilisateurs mais qu'ils s'étendront aux détenteurs de données également. Les plus gros ont de sérieuses raisons d'augmenter leur volume ; en effet, cela génère pour eux de plus grands bénéfices à un

coût tout à fait marginal. D'abord, ils disposent déjà d'une infrastructure, en termes de stockage et de traitement. Ensuite, la combinaison d'ensembles de données produit une valeur particulière. Enfin, obtenir des données sur une plateforme unique simplifie la vie des utilisateurs.

Fait plus surprenant, un nouveau type de détenteurs de données pourrait également apparaître à l'autre extrême : les personnes physiques. La valeur des données se révèle de plus en plus évidente, et cela pourrait en inciter certains à investir ce domaine en tant que détenteurs de données personnelles concernant, par exemple, leurs préférences d'achat, leurs habitudes de consommation médiatique et sans doute également les données sur leur santé.

Le fait de détenir des données personnelles confère un certain pouvoir aux consommateurs, aspect qui n'a pas été examiné auparavant. Ces personnes peuvent vouloir choisir elles-mêmes à qui accorder la licence d'utilisation de leurs données, et à quel prix. Naturellement, tout le monde ne voudra pas forcément vendre ses bits de données au plus offrant ; beaucoup seront heureux de les voir réutilisés gratuitement en contrepartie d'une meilleure qualité de service comme par exemple des recommandations de livres bien ciblées sur Amazon et une meilleure convivialité en ligne sur Pinterest, le site Web de partage de contenus et d'images axés sur les centres d'intérêt des utilisateurs. Mais pour un nombre important de consommateurs férus de technologies numériques, promouvoir et vendre leurs informations personnelles pourrait devenir aussi naturel que de bloguer, d'envoyer un tweet ou de créer un article sur Wikipédia.

Pour que cela fonctionne, l'évolution du niveau de connaissance et d'exigence des consommateurs n'est cependant pas une condition suffisante. Actuellement, ce serait trop compliqué et coûteux pour une personne d'accorder une licence d'utilisation de ses données personnelles et pour une société, de traiter directement avec elle pour les obtenir. À l'avenir, nous verrons vraisemblablement apparaître de nouvelles sociétés qui regrouperont les données provenant de nombreux consommateurs, proposeront un système simplifié d'octroi de licences et automatiseront les transactions. Si leurs coûts sont suffisamment bas et s'ils gagnent la confiance de leurs clients potentiels, il est tout à fait concevable qu'un marché des données personnelles se mette en place. Des entreprises comme Mydex en Grande-Bretagne et des groupes comme ID3, cofondé par Alex

Sandy Pentland, le grand spécialiste de l'analyse des données au MIT, œuvrent déjà à concrétiser cette vision.

Néanmoins, jusqu'à ce que ces intermédiaires soient opérationnels et que les utilisateurs de données commencent à faire appel à eux, ceux qui souhaitent devenir détenteurs de leurs propres données n'ont qu'un choix limité de possibilités. Dans l'intervalle, en attendant la mise en place de l'infrastructure et des intermédiaires, les particuliers peuvent envisager tout simplement d'en divulguer le moins possible.

Pour les sociétés de taille moyenne, les big data sont en revanche moins utiles. Les avantages d'échelle concernent les très grandes entreprises, les avantages en termes de coût et d'innovation, les petites entreprises, affirme Philip Evans du Boston Consulting Group, penseur éclairé des technologies et des affaires. Dans les secteurs traditionnels, il existe des entreprises de taille moyenne parce qu'elles combinent une certaine taille minimum nécessaire pour tirer parti des économies d'échelle avec une certaine flexibilité qui manque aux grands opérateurs. Mais dans un monde de big data, il n'existe pas d'échelle minimale que devrait atteindre une société pour réaliser ses investissements dans l'infrastructure de production. Les utilisateurs de larges quantités de données qui souhaitent réussir tout en conservant une certaine souplesse s'apercevront qu'ils n'ont plus la contrainte d'atteindre une certaine taille. Ils pourront conserver leur petite taille et néanmoins prospérer (ou bien être acquis par un géant des big data).

Les big data exercent une pression sur les entreprises de taille intermédiaire dans chaque secteur, les contraignant à choisir entre trois options : une très grande taille, une petite taille avec beaucoup de réactivité ou la mort. De nombreux secteurs traditionnels finiront par se repositionner comme des secteurs de big data, des services financiers à la fabrication en passant par le domaine pharmaceutique. Les big data n'élimineront pas systématiquement toutes les moyennes entreprises dans tous les secteurs, mais certaines, plus vulnérables à leur puissance, vont ressentir une rude pression.

Les big data sont également en passe de fragiliser les avantages concurrentiels des États. À une époque où l'industrie manufacturière a été en grande partie perdue au profit des pays en développement et où l'innovation semble ouverte à tous, les pays industrialisés conservent un avantage, celui de détenir les données et le savoir-faire pour les utiliser. La

mauvaise nouvelle est que cet avantage n'est pas durable. Comme cela s'est produit pour l'informatique et l'Internet, l'avance prise par les pays occidentaux dans le domaine des big data se réduira avec l'adoption progressive de cette technologie dans d'autres pays du monde. La bonne nouvelle pour les sociétés les plus dynamiques dans les pays développés est que les big data vont probablement accentuer les atouts et les points faibles de l'entreprise. Si une société maîtrise les big data, elle a des chances non seulement de surpasser ses concurrents mais aussi de renforcer sa position de leader.

La course est engagée. Tout comme l'algorithme de recherche de Google a besoin des traces numériques des utilisateurs pour donner de bons résultats, et tout comme le fournisseur de pièces détachées allemand avait saisi l'importance des données pour améliorer ses composants, toutes les entreprises peuvent, elles aussi, tirer profit de l'exploitation intelligente des données.

Ces perspectives réjouissantes ne vont toutefois pas sans motifs d'inquiétude. Les prévisions concernant notre monde et la place que nous y avons sont de plus en plus précises grâce à l'analyse de masses gigantesques de données mais il se pourrait bien que nous ne soyons pas vraiment préparés à faire face à l'impact de ces données sur notre vie privée et notre conception de la liberté. C'est dans un monde rare en informations que notre perception s'est formée et que nos institutions se sont établies, et non dans l'abondance. C'est donc la face sombre des big data que nous allons examiner dans le prochain chapitre.

[a.](#) Agence gouvernementale des statistiques du transport.

[b.](#) Division du Département des transports des États-Unis.

[c.](#) Service météorologique des États-Unis.

[d.](#) L'équivalent local de notre Bison futé français.

[e.](#) Loi de probabilité qui s'applique aux événements rares.

8

Risques

Durant près de quarante ans et jusqu'à la chute du mur de Berlin en 1989, le ministère de la Sécurité d'État d'Allemagne de l'Est connu sous le nom de Stasi a espionné des millions de personnes. La Stasi comptait près de cent mille employés à plein temps et sa surveillance s'exerçait partout, dans la rue, à partir de véhicules... Le courrier était ouvert, les comptes bancaires épluchés, les citoyens mis sur écoutes téléphoniques, même à la maison ! La Stasi incitait les couples, les parents et les enfants à s'espionner mutuellement, poussant à trahir ce qu'il y a de plus fondamental, la confiance entre les êtres humains. Le résultat de cette entreprise fut la constitution d'un nombre impressionnant de dossiers – notamment 39 millions de fiches et plus de 110 kilomètres de documents au moins – où était consignée la vie de gens ordinaires dans ses détails les plus intimes. L'Allemagne de l'Est a été l'un des États espions les plus totalitaires que l'on ait jamais vu.

Vingt ans après la fin de l'Allemagne de l'Est, la collecte et le stockage de données concernant chacun d'entre nous se poursuivent de plus belle. Nous sommes sous surveillance constante : lorsque nous utilisons notre carte de crédit, notre téléphone cellulaire ou notre numéro de Sécurité sociale. En 2007, les médias britanniques n'ont pas manqué de souligner l'ironie de la situation : plus de 30 caméras de surveillance étaient placées à moins de 200 mètres de l'appartement londonien où George Orwell avait écrit son roman *1984*. Bien avant l'avènement d'Internet, des sociétés spécialisées comme Equifax, Experian et Acxiom ont collecté, compilé et donné accès aux informations personnelles de centaines de millions de personnes à travers le monde. Internet a rendu le suivi plus facile, moins coûteux et plus efficace. Et les agences gouvernementales américaines aux

sigles à trois lettres et aux activités tenues secrètes ne sont pas les seules à nous espionner. Amazon surveille nos préférences en matière d'achats et Google nos habitudes de navigation, tandis que Twitter sait à quoi nous pensons. Facebook semble capter toutes ces informations, lui aussi, en même temps qu'il suit nos relations sociales. Les opérateurs mobiles savent non seulement à qui nous parlons, mais qui se trouve à proximité à ce moment-là.

Avec la promesse des big data de livrer à ceux qui les analysent de précieuses informations, tout semble indiquer une forte intensification de la collecte, du stockage et de la réutilisation de nos données personnelles par des tiers. Le volume et l'échelle des collectes de données augmenteront à pas de géant au fur et à mesure que les coûts de stockage chuteront et que les outils d'analyse deviendront encore plus performants. Si l'ère d'Internet menaçait déjà la vie privée, les big data la mettront-elles encore plus en danger ? Est-ce la face sombre des big data ?

Oui, et pas la seule. Sur ce plan aussi, un des traits essentiels des big data ressort : le changement d'échelle mène à un changement d'état. Comme nous l'expliquerons plus loin, cette transformation non seulement rend la protection de la vie privée plus difficile, mais elle présente aussi une menace totalement nouvelle : la pénalisation des intentions. C'est-à-dire la possibilité d'utiliser les prévisions des big data pour juger et punir des individus avant même qu'ils aient commis un quelconque acte répréhensible. Ce qui constitue la négation même des idées d'équité, de justice et de libre arbitre.

Outre la vie privée et l'intention, le danger menace sur un troisième front. Nous risquons d'être victimes d'une dictature des données : nous pourrions en arriver à fétichiser les informations, les résultats de nos analyses, et nous finirions par en faire un mauvais usage. Maniés avec intelligence, les ensembles de données sont un outil utile pour une prise de décision rationnelle. Détournés par les puissants de façon mal intentionnée, ils peuvent se muer en instrument de répression, qui aura pour effet au mieux de mécontenter clients et employés, au pire, de porter préjudice aux citoyens.

Les enjeux sont plus importants qu'on ne le reconnaît en général. L'absence de contrôle des big data en termes de vie privée ou des prévisions erronées font courir des dangers bien plus graves que d'être simplement ciblé par des publicités en ligne. L'histoire du xx^e siècle compte un certain

nombre de situations où les données ont servi de noirs desseins qui ont fait couler le sang. En 1943, le Bureau américain du recensement a remis aux autorités des adresses de pâtés de maisons où résidaient des Nippo-Américains (sans les noms et numéros de rue, cela pour maintenir l'apparence d'une protection de la vie privée) pour faciliter leur arrestation et leur internement. Quand les nazis ont envahi les Pays-Bas et opéré des rafles parmi les juifs, ils ont utilisé les registres d'état civil connus pour être très complets. Les numéros à cinq chiffres tatoués sur l'avant-bras des prisonniers des camps de concentration nazis correspondaient à l'origine aux numéros de cartes perforées Hollerith d'IBM ; le traitement des données a contribué au meurtre à une échelle industrielle.

En dépit de ses prouesses informationnelles, beaucoup de choses dépassaient les capacités de la Stasi. Connaître systématiquement tous les déplacements de tous les habitants ou toutes leurs conversations et leurs interlocuteurs était une entreprise laborieuse. Aujourd'hui, une bonne partie de ces informations est collectée par les opérateurs de téléphonie mobile. Pas plus que nous, l'État est-allemand ne pouvait prédire qui deviendrait dissident, tandis qu'aujourd'hui, les forces de police commencent à se servir de modèles algorithmiques pour décider où et quand patrouiller ; cela nous donne un avant-goût de ce qui nous attend. Il est ainsi aisé de comprendre que les risques inhérents aux big data sont aussi grands que les ensembles de données eux-mêmes.

Paralysie de la vie privée

Il est tentant d'établir un parallèle entre le danger que la croissance des données numériques fait peser (par extrapolation) sur la vie privée et la surveillance au cœur de *1984*, l'œuvre de contre-utopie de George Orwell. Et cependant la situation est plus complexe. Tout d'abord, on ne trouve pas de données personnelles dans tous les ensembles de données, que ce soit celles issues de capteurs installés dans les raffineries, de machines dans les ateliers d'usines, celles concernant les explosions qui surviennent dans les trous d'homme ou encore celles relatives aux conditions météorologiques dans un aéroport. Pour extraire la valeur de leurs analyses, BP et Con Edison n'ont pas besoin d'informations personnelles (et n'en recherchent d'ailleurs pas). Les analyses de ce type d'informations ne présentent pratiquement aucun risque pour la vie privée.

Néanmoins, une bonne partie des données qui sont actuellement générées contient effectivement des informations personnelles. Et les sociétés ont une multitude de motivations pour en recueillir toujours plus, les conserver plus longtemps et les réutiliser plus souvent. Même quand les données ne paraissent pas explicitement porter sur des informations personnelles, elles permettent, grâce aux procédures employées avec les big data, de remonter facilement jusqu'à la personne concernée. Ou alors il est possible d'en déduire des détails intimes sur cette personne.

Par exemple, aux États-Unis et en Europe des services publics installent des compteurs d'électricité « intelligents » qui collectent des données durant toute la journée, probablement toutes les six secondes – bien plus que le mince filet d'informations sur la consommation globale d'énergie que recueillent les compteurs classiques. Fait important, le type de consommation varie d'un appareil électrique à l'autre et cette variation constitue la « signature électrique » spécifique à chaque appareil. Ainsi, un chauffe-eau électrique est différent d'un ordinateur, qui lui-même diffère de lampes de serre pour plants de marijuana. La consommation d'énergie d'un ménage révèle donc des informations privées, que ce soit le comportement quotidien de l'habitant, son état de santé ou ses activités illégales.

La question importante n'est donc pas de savoir si le risque d'atteinte à la vie privée augmente avec les big data – on a compris que c'est bien le cas –, mais si la nature du risque se modifie. Si la menace est simplement plus grande, alors les lois et les règles qui protègent la vie privée demeurent valides à l'ère des big data ; il nous suffit de redoubler d'efforts. En revanche, si le problème change de nature, il va nous falloir trouver de nouvelles solutions.

La mauvaise nouvelle, c'est que le problème a bel et bien subi une transformation. Avec des quantités massives de données, la valeur de l'information ne se trouve plus seulement dans son objectif initial. Comme nous l'avons dit, il faut compter maintenant avec ses réutilisations.

Le rôle central assigné aux individus dans les lois sur la protection de la vie privée est ici amoindri par ce changement. Aujourd'hui, on signale à l'utilisateur, au moment de la collecte, le type d'informations recueillies et dans quel but ; chacun peut ensuite refuser ou accepter, et la collecte commencer ou non. Comme l'explique Fred Cate, expert dans le domaine de la protection de la vie privée à l'université de l'Indiana, ce concept de « notification et consentement » est devenu la clé de voûte des principes qui

ont cours en matière de protection de la vie privée à travers le monde, même s'il ne s'agit pas du seul moyen légal de recueillir et de traiter des données personnelles. (Cela donne en pratique des avis de confidentialité interminables qui sont rarement lus, et a fortiori compris – mais ceci est une autre histoire.)

Chose frappante, à l'ère des big data, la plupart des utilisations secondaires innovantes n'ont pas été envisagées au moment de la collecte initiale des données. Comment les entreprises pourraient-elles fournir une notification concernant un objectif qui n'a pas encore été défini ? Comment donner un consentement éclairé à des inconnus ? En l'absence de consentement, pour toute analyse de volumes importants de données contenant des informations personnelles, il faudrait pourtant revenir à chaque individu et lui demander son autorisation pour toute réutilisation de ses données. Pouvez-vous imaginer Google essayant de contacter des centaines de millions d'utilisateurs pour leur demander leur accord avant d'utiliser leurs anciennes requêtes dans le but de prédire la grippe ? Aucune société n'en assumerait le coût, même si la tâche était techniquement réalisable.

L'alternative, à savoir demander aux utilisateurs de donner leur accord au moment de la collecte pour toute utilisation future possible de leurs données, n'est pas utile non plus. La notion même de consentement éclairé est affaiblie par ce genre d'autorisation globale. Résultat : dans le contexte des big data, le concept éprouvé (et fiable) de notification et de consentement se retrouve souvent soit trop restrictif pour que la valeur latente des données en soit extraite, soit trop creux pour protéger un individu et sa vie privée.

Les autres moyens de protéger la vie privée sont eux aussi inefficaces. Une fois que des informations personnelles ont figuré dans un ensemble de données, même si l'on fait le choix de s'opposer à leur utilisation, cela n'empêche pas qu'il en reste une trace. Prenons le cas de Google Street View. Les voitures de la société ont collecté des images de rues et de maisons dans de nombreux pays. En Allemagne, Google s'est heurté à de fortes protestations du public et des médias. Les gens craignaient que les images de leurs maison et jardin n'aident les cambrioleurs à mieux choisir leurs cibles. Mis sous pression réglementaire, Google a accepté de laisser le choix aux propriétaires de faire opposition et dans ce cas, les images de leur maison ont été floutées. Mais ce retrait est visible sur Street View – les

maisons floutées se remarquent – et les cambrioleurs peuvent l’interpréter comme le signe d’une cible particulièrement intéressante.

Il existe une méthode technique de protection de la vie privée – l’anonymisation –, mais dans de nombreux cas, elle ne fonctionne pas bien non plus. Elle consiste à éliminer des ensembles de données tout identifiant personnel – nom, adresse, numéro de carte de crédit, date de naissance ou numéro de Sécurité sociale. Les données qui en résultent peuvent alors être analysées et communiquées sans compromettre la vie privée de qui que ce soit. Mais cette méthode ne fonctionne que dans un contexte de petits volumes de données. Les big data, qui se caractérisent par la grande quantité et la variété des informations, facilitent la réidentification. Prenons le cas de requêtes faites sur Internet à propos de classements de films, tout cela apparemment non identifiable.

En août 2006, le site AOL a diffusé pléthore de vieilles requêtes, animé de la bonne intention de proposer ainsi à l’analyse des informations éventuellement intéressantes pour ses visiteurs. L’ensemble de données, fort de 20 millions de requêtes effectuées par 657 000 utilisateurs entre le 1^{er} mars et le 31 mai de la même année, avait été soigneusement anonymisé. Les informations personnelles comme le nom d’utilisateur et l’adresse IP avaient été effacées et remplacées par des identifiants numériques uniques. L’idée était que les chercheurs pourraient faire un lien entre les requêtes d’un même individu, mais sans disposer d’informations permettant de l’identifier.

Néanmoins, en quelques jours à peine, le *New York Times*, en associant des requêtes diverses intitulées « 60 hommes célibataires », « le thé pour une bonne santé » et « des paysagistes à Lilburn, Georgie » est parvenu à identifier l’utilisateur numéro 4417749 : il s’agissait de Thelma Arnold, une veuve de soixante-deux ans de Lilburn, en Georgie. « Seigneur Dieu, mais c’est toute ma vie personnelle, déclara-t-elle au journaliste du *Times* venu frapper à sa porte. J’ignorais absolument que quelqu’un se penchait par-dessus mon épaule ! » Après le tollé général qui s’est ensuivi, le directeur de la technologie et deux autres employés d’AOL ont été limogés.

Cela n’a pas empêché Netflix, l’entreprise de location de films, de lancer en octobre 2006, à peine deux mois plus tard, le même type d’opération, appelée cette fois « Prix Netflix ». 100 millions de relevés de locations de films par près d’un demi-million d’utilisateurs ont été publiés par la société qui a offert une récompense d’un million de dollars à toute

équipe capable d'améliorer d'au moins 10 % son système de recommandation de films. Là encore, les identifiants personnels avaient été soigneusement supprimés. Et pourtant une fois de plus, une utilisatrice a été désanonymisée et donc de nouveau identifiée : une mère de famille, résidant dans le Midwest américain conservateur, qui dissimulait son homosexualité et qui, pour cette raison, a plus tard poursuivi en justice Netflix sous le pseudonyme de « Jane Doe ».

Les chercheurs de l'université du Texas à Austin ont comparé les données de Netflix avec d'autres informations publiques. Ils ont rapidement découvert comment établir les correspondances entre les classements communiqués par un utilisateur anonymisé et ceux d'un contributeur du site Web d'Internet Movie Database (IMDb)^a. De manière plus générale, la recherche a démontré qu'il suffisait simplement qu'un client de Netflix donne son classement de six films peu connus (n'entrant pas dans les 500 premiers), pour être identifié dans 84 % des cas. Et si la date de ce classement était connue, on pouvait l'identifier parmi le demi-million de clients environ figurant dans l'ensemble de données, avec un taux d'exactitude de 99 %.

Dans le cas d'AOL, les identités des utilisateurs ont été révélées par le contenu de leurs requêtes. Dans celui de Netflix, par une comparaison des données avec d'autres sources. Dans les deux exemples, les sociétés n'ont pas su évaluer à quel point les big data contribuaient à la désanonymisation, et cela pour deux raisons : nous recueillons un nombre plus grand de données et nous en combinons également un nombre plus grand.

Paul Ohm, professeur de droit à l'université du Colorado à Boulder et expert en matière de méfaits de la désanonymisation, explique qu'il n'existe pas de solution miracle. À partir d'un certain volume de données, une anonymisation parfaite est impossible, quels que soient les efforts déployés. Pis, les chercheurs ont récemment démontré que non seulement les données conventionnelles mais aussi le graphe social – les liens entre individus – peuvent faire l'objet d'une désanonymisation.

À l'ère des big data, les trois stratégies de base utilisées depuis longtemps pour protéger la vie privée – notification et consentement individuel, exclusion et anonymisation – ont perdu une bonne partie de leur efficacité. D'ores et déjà, de nombreux utilisateurs ont un sentiment de violation de leur vie privée. Qu'en sera-t-il alors lorsque l'utilisation de très grands ensembles de données sera devenue plus courante ?

Comparé à l'Allemagne de l'Est il y a un quart de siècle, la surveillance est devenue plus facile, moins coûteuse et plus puissante. La capacité de recueillir des données personnelles est souvent profondément intégrée dans les outils que nous utilisons aujourd'hui, des sites Web aux applications pour smartphones. On sait qu'au tribunal, lors des procès sur les causes d'accident, les capteurs embarqués dans la plupart des voitures capables de tout enregistrer, et notamment ce qui s'est passé quelques secondes avant l'activation de l'airbag, ont « témoigné » contre les conducteurs.

Évidemment, quand des entreprises collectent des données pour améliorer leur résultat net, il n'y a pas à craindre que leur surveillance ait les mêmes conséquences qu'une mise sur écoute par la Stasi. Nous n'irons pas en prison si Amazon découvre que nous aimons lire *Le Petit Livre rouge* du président Mao. Google ne nous exilera pas parce que nous avons fait une recherche avec le mot-clé « Bing », le moteur de recherche de Microsoft. Les sociétés peuvent être puissantes, mais elles n'ont pas le pouvoir de coercition d'un État.

Mais même si elles ne viennent pas nous arrêter en pleine nuit, des sociétés de toutes sortes engrangent des tonnes d'informations personnelles sur tous les aspects de notre vie, les communiquent à des tiers à notre insu et en font une utilisation qu'on a du mal à imaginer.

Le secteur privé n'est pas le seul à afficher ses capacités en matière de big data. Les gouvernements le font également. L'U.S. National Security Agency^b (NSA) aurait intercepté et stocké 1,7 milliard de courriels, appels téléphoniques et autres types de communications par jour, selon une enquête du *Washington Post* publiée en 2010. William Binney, ancien fonctionnaire de la NSA, estime que le gouvernement a compilé « 20 billions de transactions » entre citoyens américains et autres : qui appelle qui, qui envoie un courriel à qui, qui envoie de l'argent à qui, etc.

Pour exploiter toutes les données, les États-Unis sont en train de construire des centres de données géants comme celui de la NSA à Fort Williams, dans l'Utah, 1,2 milliard de dollars. Toutes les instances gouvernementales demandent plus d'informations qu'auparavant, et pas uniquement des organisations opaques engagées dans la lutte contre le contre-terrorisme. Quand la collecte s'étend à des informations comme les transactions financières, les dossiers de santé et les mises à jour de statuts sur Facebook, la quantité recueillie atteint des niveaux incroyables. Le

gouvernement ne peut traiter autant de données. Mais alors pourquoi les collecter ?

La réponse est à trouver dans l'évolution des méthodes de surveillance à l'ère des big data. Dans le temps, les enquêteurs fixaient des pincettes crocodile aux câbles téléphoniques pour obtenir le maximum d'informations sur un suspect. Ce qui primait était de fouiller dans la vie de la personne pour tout savoir sur elle. La méthode moderne est différente. Dans l'esprit de Google ou Facebook, l'idée nouvelle est que les individus sont la somme de leurs relations sociales, de leurs interactions en ligne et de leurs connexions avec du contenu. Pour en apprendre le plus possible sur quelqu'un, les analystes doivent ratisser large en termes d'informations : ils doivent chercher à savoir non seulement l'identité de ses relations, mais également quelles autres personnes les connaissent aussi, etc. Ce qui était techniquement difficile à réaliser autrefois n'a jamais été aussi facile qu'aujourd'hui. Et comme le gouvernement ne sait pas sur qui il voudra un jour mener une investigation, il collecte, stocke des informations ou s'en assure l'accès ; l'intention des autorités n'est pas forcément de surveiller tout le monde tout le temps, mais de pouvoir immédiatement lancer une enquête sur un individu, en cas de soupçon, plutôt que de devoir commencer à collecter les informations à partir de rien.

Le gouvernement des États-Unis n'est pas le seul à amasser des données sur les citoyens, et peut-être même pas le plus puissant. Mais cette capacité des entreprises et du gouvernement à obtenir des informations personnelles nous concernant n'est pas la seule à nous préoccuper. Il y a un nouveau problème qui se pose depuis peu avec les big data : l'utilisation des prévisions pour nous juger.

Probabilité et sanction

John Anderton dirige une unité spéciale de police à Washington, D.C. Ce matin-là, il fait irruption dans une maison de banlieue quelques instants avant qu'Howard Marks, dans un état de rage folle, ne plonge une paire de ciseaux dans la poitrine de sa femme, qu'il a trouvée au lit avec un autre homme. Pour Anderton, ce n'est qu'un jour parmi d'autres passés à prévenir des crimes capitaux. « Par ordre de la Division précrime du district de Columbia, récite-t-il, je vous arrête pour le meurtre futur de Sarah Marks, que vous alliez commettre aujourd'hui... »

D'autres policiers commencent à immobiliser Marks, qui hurle : « Je n'ai rien *fait* ! »

La scène d'ouverture du film *Minority report* (*Rapport minoritaire*) dépeint une société dans laquelle les prévisions semblent si exactes que la police arrête des individus pour des crimes qui n'ont pas encore été perpétrés. On emprisonne l'individu non pas pour ses actes, mais pour ce qu'on anticipe de ses actions, même si dans la réalité, il ne commettra jamais le crime en question. Dans le film, ce sont les visions de trois voyants, et non l'analyse de données, qui sont à la source d'un maintien de l'ordre fondé sur la précognition et la prévention. Mais une analyse non contrôlée des données menace de nous mener vers ce futur inquiétant décrit dans le film où une personne est jugée coupable sur la base de prédictions concernant son comportement ultérieur.

Nous en voyons déjà les prémices. En Amérique, les Commissions de libération conditionnelle dans plus de la moitié des États se servent de prévisions fondées sur l'analyse de données pour décider de remettre un prisonnier en liberté ou de le maintenir en incarcération. La méthode dite de « police préventive » est de plus en plus pratiquée – depuis des postes de police de Los Angeles jusqu'à des villes comme Richmond, en Virginie – : l'analyse d'ensembles de données permet de sélectionner les rues, mais aussi les groupes et les individus qui devront être particulièrement surveillés, simplement parce qu'un algorithme a indiqué qu'il y avait de fortes probabilités qu'un crime soit commis.

Dans la ville de Memphis, au Tennessee, un programme du nom de Blue CRUSH (acronyme de Crime Reduction Utilizing Statistical History⁴) fournit aux officiers de police des indications relativement précises et intéressantes pour eux en termes de localisation (quelques pâtés de maisons) et de période (quelques heures tel ou tel jour de la semaine). Le système permet nettement aux maigres effectifs de la police de mieux cibler les missions. Depuis sa mise en place en 2006, le nombre de crimes graves avec atteinte à la propriété et d'infractions avec violence a chuté d'un quart, d'après les mesures utilisées (mais bien sûr, cela ne dit rien sur le lien de cause à effet ; rien n'indique que la baisse doive être attribuée à Blue CRUSH).

À Richmond, en Virginie, la police fait des corrélations entre des données concernant les crimes et d'autres types d'informations telles que les dates de paye des grandes sociétés de la ville ou celles de concerts ou

d'événements sportifs. Une démarche qui a confirmé et parfois précisé les soupçons que la police avait sur les tendances en matière de criminalité. Par exemple, il lui semblait que les crimes violents augmentaient fortement suite aux expositions d'armes à feu ; l'analyse de données a corroboré leur impression, à un détail près : le taux de criminalité a augmenté deux semaines plus tard et non pas immédiatement après l'événement.

Le but de ces systèmes est de prévenir les crimes en étant capable de préciser jusqu'à la personne qui pourrait les commettre. Il s'agit là d'un nouvel objectif de l'utilisation des big data : empêcher le crime de se produire.

Un projet de recherche relevant du Département de la sécurité intérieure des États-Unis appelé FAST (Future Attribute Screening Technology^d) a pour but d'identifier des terroristes potentiels en surveillant les signes vitaux, le langage corporel et d'autres caractères de la physiologie chez les individus. L'idée est qu'en surveillant le comportement, il serait possible de détecter l'intention de nuire. D'après le DHS, les tests montrent que le système se révèle exact dans 70 % des cas. (L'explication n'est pas tout à fait claire ; les sujets soumis à cette recherche avaient-ils pour instruction de faire semblant d'être des terroristes pour qu'on puisse vérifier que leur « intention de faire du mal » serait détectée ?) Ces systèmes ont beau paraître embryonnaires, le fait est que la police les prend très au sérieux.

Empêcher qu'un crime soit perpétré a tout l'air d'une perspective intéressante. Ne serait-il pas plus utile de faire la prévention des infractions avant qu'elles n'aient été commises plutôt que d'en punir les auteurs après coup ? La prévention de la criminalité ne profiterait-elle pas à la société dans son ensemble et pas seulement à ceux qui auraient pu en être les victimes directes ?

Mais c'est là s'engager sur un chemin périlleux. Si par le biais des big data, il nous devenait possible de connaître à l'avance l'auteur d'un crime, nous pourrions ne plus nous contenter de simplement prévenir ce crime ; nous serions susceptibles de vouloir aussi en punir l'auteur probable. Cela tombe sous le sens. Si nous nous contentons simplement d'une intervention qui empêche l'acte illicite, son auteur présumé pourrait bien être tenté de le commettre à nouveau et en toute impunité. En revanche, si les big data sont utilisées pour le déclarer responsable de ses actes (futurs), cela peut devenir fortement dissuasif.

Cette sanction fondée sur la prévision nous donne l'impression d'être une amélioration par rapport aux pratiques que nous acceptons depuis des années. Prévenir les comportements malsains, dangereux ou à risque est l'une des clés de voûte de la société moderne. Nous avons rendu les choses difficiles pour les fumeurs dans le but de prévenir le cancer du poumon ; nous imposons le port de ceintures de sécurité pour éviter les morts sur la route ; nous ne laissons pas monter à bord un passager transportant des armes pour éviter les détournements d'avion. Que notre liberté soit restreinte par de telles mesures de prévention nous semble généralement un faible prix à payer pour éviter des préjudices plus graves.

Dans de nombreux contextes, l'analyse de données est déjà employée au nom de la prévention. Les cohortes des études statistiques rassemblent ceux qui ont une caractéristique en commun. Les tableaux actuariels indiquent que les hommes âgés de plus de 50 ans sont sujets au cancer de la prostate, et les membres de ce groupe peuvent se voir obligés de payer plus cher pour une assurance maladie quand bien même ils ne seraient jamais atteints d'un cancer de la prostate. Les lycéens qui obtiennent de bonnes notes, en tant que groupe, sont moins susceptibles d'être impliqués dans des accidents de voiture – et certains de leurs camarades d'un moins bon niveau doivent donc payer des primes d'assurance plus élevées. Selon leurs caractéristiques, certains individus sont soumis à des vérifications supplémentaires quand ils passent le contrôle de sécurité dans un aéroport.

C'est l'idée qui sous-tend le « profilage » dans le monde des petits volumes de données actuel : trouver un lien que les données ont en commun, définir un groupe de gens à qui ce lien s'applique et procéder ensuite à son examen plus attentif. Une règle que l'on peut généraliser à chacun des membres de ce groupe. Le « profilage », bien sûr, est un terme lourd de sens, et la méthode présente de grands inconvénients. Mal utilisée, elle peut conduire non seulement à la discrimination à l'égard de certains groupes mais aussi à désigner des « coupables par association ».

Par contre, les prévisions des big data concernant les individus ne sont pas du même ordre. Actuellement, les prévisions s'appliquant à un comportement hypothétique – dans un contexte de primes d'assurance ou de pointages de crédit – reposent généralement sur une poignée de critères (antécédents médicaux ou remboursement d'un prêt). Avec l'analyse de grandes quantités de données sans recherche de causalité, nous nous

contentons souvent d'identifier les variables prédictives les plus adéquates, elles-mêmes issues d'une foultitude d'informations.

Point très important, lorsque nous utilisons les big data, nous espérons identifier des individus spécifiques plutôt que des groupes ; cela nous permet d'échapper à l'inconvénient du profilage dont la prévision fait de chaque suspect un coupable par association. Dans un monde de big data, un individu portant un patronyme arabe, qui a payé en liquide un billet aller simple en première classe, ne sera plus soumis à un deuxième contrôle dans un aéroport si d'autres données qui lui sont propres réduisent fortement la probabilité qu'il soit un terroriste. Avec les big data, il devient possible d'échapper aux contraintes liées aux identités de groupes, et de les remplacer par des prévisions d'une granularité^e bien plus adaptée à l'individu.

Ce qu'on peut attendre des big data, c'est que nous continuions à procéder comme nous l'avons toujours fait – méthode du profilage – mais en affinant le résultat pour qu'il soit moins discriminatoire et plus individualisé. Cela paraît acceptable si le but demeure simplement de prévenir des actions indésirables. En revanche, cela devient très dangereux si les prévisions issues d'ensembles de données nous servent à décider de la culpabilité d'un individu et de la sanction à exécuter pour un acte qu'il n'a pas encore commis.

L'idée même de pénaliser quelqu'un à cause de ses intentions est absolument abjecte. Accuser une personne pour un acte qu'elle pourrait commettre à l'avenir, c'est nier le fondement même de la justice : à savoir qu'il faut avoir effectivement commis quelque chose avant d'en être tenu pour responsable. Après tout, avoir de mauvaises pensées n'a rien d'illégal ; ce qui l'est, c'est le passage à l'acte. Le lien entre responsabilité individuelle et choix délibéré de nos actes est un principe fondamental dans notre société. Si quelqu'un est contraint sous la menace d'une arme d'ouvrir le coffre-fort de la société, il n'a pas le choix et n'est donc pas tenu pour responsable.

Si les prévisions des big data étaient parfaites, si les algorithmes pouvaient prédire notre avenir de façon impeccable, nous n'aurions plus le choix de nos actes. Nous nous comporterions exactement suivant les prédictions. Si la perfection dans la prévision existait, cela reviendrait à nier toute volonté humaine ainsi que notre capacité à vivre notre vie librement.

Ironie de la chose, en nous privant de la capacité de choix, elles nous exonéreraient également de toute responsabilité.

Il est évident que la prévision parfaite n'existe pas. L'analyse de big data pourra seulement prédire qu'il existe telle ou telle probabilité que tel ou tel individu adopte à l'avenir tel ou tel comportement. Prenons l'exemple de la recherche que mène Richard Berk, professeur de statistiques et de criminologie à l'université de Pennsylvanie. Selon sa méthode, affirme-t-il, il est possible de prédire qu'une personne libérée sur parole sera impliquée dans un crime (qu'il tuera ou sera tué). Comme données d'entrée, il utilise de nombreuses variables spécifiques, notamment le motif de l'incarcération et la date de la première infraction, mais aussi les données démographiques comme l'âge et le sexe. Berk laisse entendre qu'il peut prévoir qu'un crime sera commis dans le futur par des individus en liberté conditionnelle avec un taux de probabilité de 75 % minimum. Pas mauvais comme score, mais cela signifie que si les commissions de libération conditionnelle se reposaient sur l'analyse de Berk, elles se tromperaient au moins une fois sur quatre.

On voit que le fond du problème, si on s'appuie sur de telles prévisions, n'est pas de savoir à quels risques la société s'expose. Mais qu'avec un tel système, nous allons punir des gens *avant* qu'ils n'aient commis un méfait. En intervenant de manière préventive (par exemple en leur refusant la libération conditionnelle parce que la probabilité est forte qu'ils commettent un meurtre, selon ce qu'indiquent les prévisions), nous ne saurons jamais si ces personnes auraient effectivement commis le crime prédit. Nous ne laissons pas le destin jouer son rôle et, pourtant, nous tenons des individus responsables d'actes hypothétiques « annoncés » par des prévisions qui ne pourront jamais être vérifiées.

Cela va à l'encontre de l'idée même de présomption d'innocence, principe sur lequel notre système juridique, ainsi que notre sens de l'équité, sont fondés. Et si nous tenons des personnes pour responsables d'actes futurs « prédits », mais qui pourraient ne jamais être commis, nous nions également aux êtres humains la capacité de faire un choix moral.

Le point important ici ne se limite *pas* simplement à l'aspect policier des choses. Le danger va bien au-delà de la justice pénale ; il couvre tous les aspects de la société, tous les cas où le jugement humain s'appuie sur des prévisions issues de big data pour décider si quelqu'un est coupable ou pas d'un acte futur – cela peut aller de la décision d'une société de licencier

un employé jusqu'à la demande de divorce d'un conjoint en passant par le refus d'un médecin d'opérer un patient.

Sans doute, avec un tel système, la société serait plus en sécurité ou fonctionnerait mieux, mais une part essentielle de notre essence humaine – notre capacité à faire le choix de nos actes et à en être tenus responsables – serait détruite. Les big data seraient alors devenues un outil d'uniformisation des décisions humaines et d'abandon du libre arbitre dans notre société.

Certes, les big data présentent de multiples avantages. Le défaut qui les transforme en arme de déshumanisation n'est pas inhérent à leur nature, mais tient à l'utilisation que nous voulons faire des prévisions. Le nœud du problème réside dans la contradiction qui consiste à prendre des décisions d'ordre causal en matière de responsabilité individuelle (considérer les gens coupables d'actes avant qu'ils ne les aient commis en utilisant des prévisions de big data) et le fait que les prévisions sont fondées sur des corrélations.

Les big data sont utiles pour comprendre les risques présents et futurs, et pour adapter nos actions en conséquence. Leurs prévisions aident les patients et les assureurs, les prêteurs et les consommateurs. Mais elles ne nous apportent rien sur le plan de la causalité. En revanche, pour déclarer quelqu'un « coupable » – dans le sens d'une culpabilité individuelle – il faut que celui-ci ait choisi d'accomplir un acte particulier. Sa décision doit avoir été la cause de l'acte qui a suivi. Précisément parce que les big data sont fondées sur des corrélations, elles constituent un outil totalement inadéquat pour juger des actes qui relèvent d'une causalité et donc pour désigner un individu comme étant coupable.

Le problème est que les êtres humains sont disposés à voir le monde à travers le prisme de la cause et de l'effet. Voilà pourquoi les big data risquent en permanence d'être utilisées de façon abusive et d'être prises pour la solution idéale, qui nous permet d'être plus efficace dans nos jugements et de mieux décider de la culpabilité de tel ou tel.

C'est là que réside essentiellement le risque de dérive – celui qui conduit tout droit vers la société décrite dans *Minority Report*, un monde d'où le choix individuel et le libre arbitre ont été éliminés. Un monde dans lequel notre sens moral personnel a été remplacé par des algorithmes prédictifs, un monde où les individus sont exposés à la violence d'un diktat collectif. Si on les utilise de cette manière, la menace que les big data

feraient alors peser sur nous, peut-être même au sens propre, ce serait de nous retrouver emprisonnés dans les probabilités.

La dictature des données

Les big data constituent une atteinte à la vie privée et une menace pour la liberté. Mais avec elles, un autre problème très ancien s'aggrave, celui qui consiste à s'appuyer sur les chiffres alors qu'ils sont bien plus faillibles que nous ne le pensons. Rien ne met plus en évidence les conséquences d'un dérapage en matière d'analyse des données que l'histoire de Robert McNamara.

McNamara était un homme férù de statistiques. Nommé Secrétaire américain à la Défense au moment où des tensions commençaient à se manifester au Vietnam, au début des années 1960, il exigeait d'obtenir des données sur tout ce qui était possible. Seule l'utilisation rigoureuse de statistiques, pensait-il, pouvait permettre aux décideurs d'appréhender une situation complexe et de faire les bons choix. Le monde, selon lui, n'était qu'une masse d'informations désorganisées qui, si elles étaient délimitées, encodées, démarquées et quantifiées, pourraient être maîtrisées par l'homme et se plieraient à sa volonté. McNamara était en quête de la Vérité, et cette Vérité se trouvait dans les données. Parmi les chiffres qui lui étaient fournis figurait le « décompte des morts ».

La passion des chiffres est venue à McNamara quand il était étudiant à la Harvard Business School avant d'y devenir le plus jeune professeur adjoint à l'âge de vingt-quatre ans. Il a fait preuve de la même rigueur au cours de la Seconde Guerre mondiale au sein de l'équipe d'élite du Pentagone réunie sous le nom de « Contrôle statistique », qui a fait entrer dans l'une des plus grandes bureaucraties du monde le processus de prise de décision fondé sur les données. Avant cela, l'armée fonctionnait à l'aveuglette. Elle ignorait, par exemple, le type, la quantité de pièces détachées par avion et où elles se trouvaient. Les données sont venues remettre bon ordre dans tout cela. Le simple fait d'avoir amélioré l'efficacité des achats d'armement a fait économiser 3,6 milliards de dollars en 1943. Dans la guerre moderne, l'allocation efficace de ressources était essentielle ; le travail de l'équipe fut une réussite spectaculaire.

À la fin de la guerre, le groupe a décidé de poursuivre ensemble le travail et de proposer ses talents aux entreprises américaines. La société

Ford Motor était en difficulté et c'est un Henry Ford II désespéré qui leur a passé les rênes. De la même manière qu'ils ignoraient tout de l'armée lorsqu'ils avaient contribué à la victoire, ils ne connaissaient rien à la fabrication de voitures. Cela n'empêcha pas ceux qu'on appelait « les Petits Génies » de redresser la société.

McNamara a rapidement gravi les échelons, sortant de sa manche un point de données pour chaque situation. Des directeurs d'usine stressés fournissaient les chiffres – corrects ou pas – qu'il réclamait. Lorsque les instructions ont été données d'utiliser tous les stocks liés à un modèle de voiture existant avant d'entamer la production d'un nouveau modèle, des responsables de chaînes de production exaspérés ont tout simplement jeté les stocks de pièces détachées excédentaires dans une rivière voisine. Les hauts responsables au siège ont manifesté leur approbation quand les contremaîtres ont renvoyé des chiffres confirmant que les ordres avaient été suivis. Pendant ce temps à l'usine, la blague qui courait était qu'on pouvait carrément marcher sur l'eau – sur un tapis de pièces détachées rouillées de voitures de l'année 1950 et de l'année 1951.

McNamara a incarné le parfait exemple du gestionnaire de la moitié du xx^e siècle, du dirigeant hyper-rationnel qui se fiait plus aux nombres qu'aux sentiments et qui pouvait mettre ses compétences au service de n'importe quel secteur d'activité de son choix. En 1960, il fut nommé président de la société Ford, poste qu'il n'a occupé que quelques semaines avant d'être nommé secrétaire à la Défense par le président Kennedy.

Avec l'escalade de la guerre et l'envoi par les États-Unis de nouvelles troupes, il apparut clairement qu'au Vietnam, on assistait à un bras de fer et non à une guerre pour le territoire. La stratégie des États-Unis était le pilonnage des soldats vietcong de manière à contraindre ce front de libération national à venir à la table des négociations. Le moyen de mesurer l'avancée était donc de dénombrer les ennemis tués. Le décompte des morts était publié quotidiennement dans les journaux. Les partisans de la guerre le considéraient comme une preuve de progrès ; pour les opposants, c'était la preuve de son immoralité. Ce décompte fut le point de données qui définit une époque.

En 1977, deux ans après que le dernier hélicoptère eut décollé du toit de l'ambassade des États-Unis à Saigon, un général de l'armée à la retraite, Douglas Kinnard, a publié une enquête qui a fait date sur les points de vue des généraux. Intitulé *The War Managers*, le livre révélait le borbier de la

quantification. À peine 2 % des généraux américains considéraient que le décompte des morts était une méthode valide pour mesurer une progression. Près des deux tiers déclaraient que les chiffres étaient souvent gonflés. « Un faux – totalement sans valeur », écrivit un général dans ses commentaires. « Mensonges flagrants bien souvent », écrivit un autre. « Les chiffres ont été nettement exagérés par de nombreuses unités principalement en raison de l'incroyable intérêt manifesté par des gens comme McNamara », déclarait un troisième.

Tout comme les ouvriers de l'usine Ford qui jetaient des pièces détachées dans la rivière, les jeunes officiers fournissaient parfois à leurs supérieurs des chiffres impressionnants pour conserver leurs galons ou bien augmenter leurs chances de promotion – et racontaient aux haut gradés ce qu'ils voulaient entendre. McNamara et son entourage se fiaient aux chiffres, les fétichisaient. Les cheveux impeccablement peignés vers l'arrière et la cravate parfaitement nouée, McNamara pensait qu'il ne pouvait appréhender ce qui se passait sur le terrain qu'en fixant un tableau de lignes et de colonnes, de calculs et de graphiques bien ordonnés, dont la maîtrise paraissait « resserrer l'intervalle » entre Dieu et lui.

L'utilisation, l'abus et l'exploitation à mauvais escient des données par l'armée américaine durant la guerre du Vietnam donnent une leçon inquiétante sur la limite qu'il faut reconnaître aux informations à l'époque des petits volumes de données ; leçon dont il faut tenir compte à un moment où le monde s'engouffre dans l'ère des big data. La qualité des données sous-jacentes peut être mauvaise. Les données peuvent être biaisées. Elles peuvent être mal analysées ou utilisées de manière erronée. Plus fâcheux encore, les données peuvent ne pas saisir ce qu'elles sont censées quantifier.

Nous sommes plus vulnérables que nous ne le pensons à la « dictature des données » – autrement dit au fait de laisser les données nous régenter. Ce qui peut se révéler aussi nuisible que bénéfique. Ce qui nous menace, c'est de nous retrouver un jour aveuglément liés au résultat de nos analyses même si nous suspectons à raison que quelque chose cloche. Ou alors de se muer en obsédés de la collecte de faits et de chiffres pour l'amour des données. Ou encore d'attribuer aux données un degré de vérité qu'elles ne méritent pas.

Du fait que la mise en données concerne de plus en plus d'aspects de la vie, la solution que les dirigeants et les hommes d'affaires commencent à

adopter c'est d'en demander encore plus. « Nous croyons en Dieu – tous les autres, apportez-nous des données » ; tel est le mantra du gestionnaire moderne, dont l'écho parvient jusqu'aux bureaux-cabines de la Silicon Valley, jusque dans les ateliers d'usines et les couloirs des agences gouvernementales. L'idée n'est pas mauvaise, mais on peut aisément être induit en erreur par les données.

L'éducation semble battre de l'aile ? Sortez la panoplie de tests normalisés pour mesurer la performance et pour pénaliser les enseignants ou les écoles qui, d'après cette mesure, ne sont pas à la hauteur. Il reste à savoir si les tests réussiront à refléter réellement les capacités des écoliers, la qualité de l'enseignement ou la nécessité de disposer d'effectifs dotés d'un esprit créatif et d'une capacité d'adaptation. Mais avec les données, pas de place pour cette question.

Vous voulez prévenir le terrorisme ? Constituez des piles de listes de surveillance et de listes de passagers interdits de vol de manière à contrôler les airs. Mais il est difficile d'affirmer que ces ensembles de données offriront effectivement la protection qu'ils promettent. Au cours d'un incident notoire, le défunt Ted Kennedy, alors sénateur du Massachusetts, a trouvé à ses dépens son nom inscrit sur une liste de passagers interdits de vol ; il a été arrêté et interrogé pour la simple raison qu'il portait le même nom qu'une personne figurant dans la base de données.

Les professionnels du monde des données ont une expression fleurie pour décrire certains de ces problèmes : « de la m... à l'entrée, de la m... à la sortie »^f. Dans certains cas, l'explication se trouve dans la qualité des informations de départ. Mais souvent, la cause en est une mauvaise utilisation de l'analyse qui en est faite. Avec les big data, ces problèmes peuvent se produire plus fréquemment ou bien avoir des conséquences plus lourdes.

Chez Google, comme nous l'avons déjà montré dans de nombreux exemples, tout fonctionne selon les données. Cette stratégie a visiblement contribué à la majeure partie de son succès, mais il arrive aussi que, de temps à autre, la société se leurre. Ses cofondateurs, Larry Page et Sergey Brin, ont longtemps tenu à connaître les notes obtenues au SAT^g par tous les candidats à un poste ainsi que leurs moyennes au diplôme d'études supérieures. Selon ce raisonnement, la première donnait la mesure du potentiel et la deuxième, celle des réalisations. Lors de leur recrutement, des cadres quadras chevronnés étaient complètement stupéfaits de se voir

demander les notes qu'ils avaient eues à l'université. La société continua même à exiger ces notes bien après que les études effectuées en interne eurent démontré l'absence de corrélation entre elles et la performance professionnelle.

Google devrait être plus avisé et ne pas céder à la séduction trompeuse des données. La mesure laisse peu de place aux changements qui adviennent au cours de la vie. Elle prend surtout en compte nos connaissances théoriques et néglige notre savoir réel. Elle ne reflétera pas forcément les qualifications des diplômés en sciences humaines, domaine où le savoir-faire pourrait être moins aisément quantifiable que dans les sciences exactes et l'ingénierie. Que Google soit obsédé par ces informations pour son recrutement semble particulièrement incongru sachant que les fondateurs de la société sont des produits des écoles Montessori, dont l'enseignement met l'accent sur l'apprentissage et non sur les notes. Et la société répète les erreurs commises dans le passé par de grandes entreprises de technologie qui plaçaient le curriculum vitae des candidats largement au-dessus de leurs compétences réelles. Larry et Sergey, qui avaient laissé tomber leurs études de doctorat, auraient-ils eu la moindre chance de décrocher un poste de directeur au sein de la légendaire institution Bell Labs^h ? Selon les normes de Google, ni Bill Gates ni Mark Zuckerberg ni Steve Jobs n'auraient été recrutés, du fait qu'ils ne possédaient aucun diplôme universitaire.

Le crédit accordé aux données par Google semble parfois excessif. Marissa Mayer, quand elle y occupait un poste de cadre dirigeant, a un jour ordonné au personnel de tester 41 nuances de bleu pour voir lesquelles étaient les plus utilisées, le but de l'opération étant de choisir la couleur d'une barre d'outils pour le site. Le respect manifesté par Google à l'égard des données a été poussé au point de déclencher une révolte.

En 2009, le concepteur principal de Google, Douglas Bowman, a quitté la société, excédé de constater que tout était quantifié en permanence. « J'ai eu récemment une discussion sur la question de savoir si une bordure devait avoir une largeur de 3, 4 ou 5 pixels, et on m'a demandé de justifier mon choix. Il m'est impossible de fonctionner dans un tel contexte, a-t-il écrit sur son blog en annonçant sa démission. Quand une société dispose d'un grand nombre d'ingénieurs dans son personnel, elle a recours à l'ingénierie pour résoudre un problème. Réduisez chaque décision à un seul problème de logique et toutes les données finissent par se transformer en une béquille

sans laquelle la moindre décision ne peut être prise et la société s'en retrouver paralysée. »

Le génie ne dépend pas des données. Steve Jobs a sans doute tenu compte des rapports de terrain pour ne cesser d'améliorer l'ordinateur portable Mac au fil des ans, mais il s'est servi de son intuition, et non de données, pour lancer l'iPod, l'iPhone et l'iPad. Il s'est fié à son sixième sens. « Ce n'est pas le boulot des consommateurs de savoir ce qu'ils veulent » : c'est la formule devenue célèbre lancée à un journaliste lors d'un entretien où il expliquait qu'Apple n'avait réalisé aucune étude de marché avant de commercialiser l'iPad.

Dans l'ouvrage *Seeing Like a State*, l'anthropologue James Scott de l'université Yale expose la façon dont les gouvernements, avec leur obsession pour la quantification et les données, finissent par rendre la vie des citoyens misérable au lieu de l'améliorer. Ils réorganisent les communautés à partir de cartes au lieu d'aller enquêter sur le terrain. Ils décident d'opter pour un mode d'agriculture collectiviste à partir de longs tableaux de données concernant les récoltes, sans avoir la moindre notion sur l'agriculture. Ils s'emparent de la façon naturelle et imparfaite dont les humains interagissent depuis la nuit des temps et l'adaptent à leurs besoins, parfois simplement pour satisfaire un désir d'ordre quantifiable. L'utilisation des données, selon Scott, a souvent pour fonction de renforcer le pouvoir des puissants.

C'est ce qu'on peut littéralement appeler la dictature des données. C'est cette même arrogance qui a entraîné dans l'escalade de la guerre du Vietnam les États-Unis, qui n'ont pas fondé leurs décisions sur des paramètres significatifs mais plutôt sur le décompte des pertes humaines. « Il est vrai qu'il n'est pas possible de réduire toute la complexité des situations humaines imaginables à quelques simples lignes dans un graphique, à des pourcentages dans un tableau ou à des chiffres dans un bilan, a déclaré McNamara dans un discours prononcé en 1967, au moment où les manifestations s'amplifiaient dans le pays. Mais toute la réalité peut être abordée de manière rationnelle. Et ne pas quantifier ce qui peut l'être revient à se limiter dans l'usage que l'on fait de la raison. » Il faudrait seulement que les bonnes données soient utilisées de façon judicieuse, au lieu d'être respectées pour ce qu'elles sont en tant que telles.

Robert Strange McNamara a été président de la Banque mondiale durant les années 1970, pour ensuite se présenter telle une colombe dans les

années 1980. Il s'est mis à dénoncer ouvertement le recours aux armes nucléaires et a viré défenseur de la protection de l'environnement. Encore plus tard, il a écrit ses Mémoires, *Avec le recul*¹, dans lesquels, ses idées ayant évolué, il faisait la critique des arguments en faveur de la guerre et de ses propres décisions en tant que secrétaire à la Défense. « Nous nous sommes trompés, nous nous sommes terriblement trompés », écrivit-il. Mais ce propos concernait la stratégie de guerre globale. Sur la question des données et du décompte des pertes humaines en particulier, il n'a exprimé aucun repentir. Il a admis qu'un bon nombre des statistiques utilisées étaient « trompeuses ou erronées ». « Mais s'il y a des choses que l'on peut compter, alors il faut les compter. Les pertes humaines en font partie... » McNamara est décédé en 2009 à l'âge de quatre-vingt-treize ans ; c'était un homme intelligent, mais dénué de sagesse.

Les big data peuvent nous conduire à commettre le péché de McNamara : nous focaliser tellement sur les données, être tellement obsédé par le pouvoir et les promesses qu'elles offrent, que l'on aurait du mal à en évaluer les limites. Pour avoir un aperçu de l'équivalent en termes de big data du décompte de pertes humaines, il nous suffit de revenir à l'outil Google Suivi de la grippe. Imaginons la situation, pas tout à fait invraisemblable, où une souche mortelle de grippe ravage le pays. Le corps médical aimerait bien pouvoir disposer d'un moyen de prévoir en temps réel les zones à plus haut risque, grâce à des requêtes sur Internet. Il saurait alors où intervenir pour porter secours aux malades.

Mais imaginons la situation suivante : en pleine crise politique, les dirigeants avancent qu'on ne peut pas se contenter du simple fait de savoir où la maladie va probablement s'aggraver et de tenter de la prévenir. Ils exigent l'instauration d'une quarantaine générale – mais pas pour tous les habitants des régions concernées. Cela serait inutile et aurait une portée excessive. Les big data nous donnent le moyen d'être plus sélectif. La décision de quarantaine ne s'appliquerait donc qu'aux internautes dont les recherches seraient en plus forte corrélation avec le fait d'avoir la grippe. Ainsi, nous avons des données qui permettent de repérer les personnes qu'il faudra regrouper dans les centres de quarantaine. (Les agents fédéraux, en effet, disposent des listes d'adresses de protocole Internet et d'informations GPS provenant de téléphones mobiles.)

Mais tout raisonnable que puisse paraître un tel scénario, il est tout bonnement erroné. Les corrélations n'impliquent pas de lien de causalité.

Ces personnes peuvent avoir la grippe ou pas. Il faudrait leur faire subir des tests. Ils seraient les prisonniers d'une prédiction, mais plus grave encore, ils seraient les victimes d'une conception des données révélatrice d'une incapacité à saisir le sens véritable des informations. L'étude réelle de Google Suivi de la grippe a permis de constater une *corrélation* entre certaines requêtes et l'épidémie – mais cette corrélation peut être le fait de circonstances ; il peut s'agir de personnes en bonne santé qui ont entendu un de leurs collègues éternuer au bureau et qui sont allées rechercher sur Internet des renseignements sur les moyens de se protéger, et non pas de personnes déjà malades.

La face sombre des big data

Comme cela a déjà été évoqué, les big data ouvrent la porte à une surveillance accrue de notre vie en même temps qu'elles rendent en grande partie obsolètes certains des moyens juridiques de protection de la vie privée. Elles privent également d'effet les procédés techniques connus pour préserver l'anonymat. Fait tout aussi préoccupant, les prévisions issues des big data peuvent être utilisées pour sanctionner certains individus pour leurs intentions, et non pour leurs actes. Cela constitue un déni du libre arbitre et une atteinte à la dignité humaine.

En parallèle, il existe un risque réel que l'on soit incité par les avantages des big data à employer des techniques à des fins non adaptées, ou à suivre aveuglément les résultats des analyses. Avec l'amélioration constante des prévisions issues des big data, il n'en sera que plus tentant de les utiliser, et leurs énormes possibilités iront un peu plus alimenter la fascination qu'elles exercent. C'est ce qui s'est produit pour McNamara et c'est la leçon à tirer de son histoire.

Nous devons nous garder d'avoir une trop grande confiance dans les données si nous ne voulons pas répéter l'erreur d'Icare : subjugué par sa propre prouesse technique, il n'a pas su la maîtriser, a volé trop haut et été précipité dans la mer. Dans le chapitre suivant, nous examinerons les moyens de contrôler les big data, pour éviter de passer sous leur contrôle.

[a.](#) Base de données cinématographiques d'Internet.

[b.](#) Agence américaine de sécurité nationale. Depuis la parution initiale du livre, cette agence est devenue célèbre même auprès du grand public, après les révélations de Edward Snowden et sa mise en cause au niveau international au sujet de son

utilisation abusive des données dans de nombreux articles de presse, notamment les écoutes de chefs d'État. (N.d.É.)

[c.](#) Réduction de la criminalité à l'aide de l'histoire statistique.

[d.](#) Technologie de surveillance des attributs futurs.

[e.](#) La granularité indique la taille du plus petit élément d'un système. On ne peut aller au-dessous et découper plus précisément l'information.

[f.](#) Ou en langage plus politiquement correct « à données inexactes, résultats erronés ».

[g.](#) Tests d'entrée dans l'enseignement supérieur.

[h.](#) Bell Telephone Laboratories : Laboratoires Bell fondés en 1925.

[i.](#) Robert S. McNamara, *Avec le recul. La tragédie du Vietnam et ses leçons*, Paris, Le Seuil, 1998.

9

Contrôle

Notre mode de production d'informations a changé, de même que nos interactions avec elles. Du coup, nos règles doivent changer, et les valeurs à protéger aussi. Nous en avons un bel exemple dans le passé, avec l'invention de l'imprimerie qui a déclenché un déluge de données.

Avant l'arrivée des caractères mobiles inventés vers 1450 par Johannes Gutenberg, la diffusion des idées en Occident se limitait principalement aux cercles de relations personnelles. Les livres étaient essentiellement confinés dans les bibliothèques des monastères et étroitement surveillés par des moines agissant pour le compte de l'Église catholique afin de protéger et de préserver sa domination. En dehors de l'Église, les livres étaient extrêmement rares. Certaines universités n'avaient réussi à recueillir que quelques dizaines de livres ou au mieux deux cents. Au début du xv^e siècle, l'université de Cambridge comptait à peine 122 volumes.

Quelques décennies après l'invention de Gutenberg, il y avait des répliques de sa presse à imprimer dans toute l'Europe, ce qui permit la production en masse de livres et de brochures. Lorsque Martin Luther traduisit la Bible latine en allemand, langue vernaculaire, la population eut soudainement une forte motivation pour apprendre à lire et à écrire : en sachant soi-même lire la Bible, on pouvait se passer des prêtres pour apprendre la parole de Dieu. La Bible devint d'ailleurs le livre le plus vendu. Mais une fois alphabétisés, les gens se mirent à lire. Certains décidèrent même d'écrire. En l'espace de moins d'une génération, le mince filet qu'était le flux d'informations se transforma en véritable torrent.

Ce changement spectaculaire a également préparé le terrain pour de nouvelles règles destinées à régir l'explosion d'informations déclenchée par les caractères mobiles. Quand l'État séculier a consolidé son pouvoir, il a

instauré un système de censure et d'autorisation afin de mettre des limites et d'imposer un contrôle sur les textes imprimés. Le droit d'auteur a été mis en place pour offrir aux auteurs des encouragements juridiques et économiques à la création. Plus tard, les intellectuels ont, eux, insisté pour l'établissement de règles destinées à protéger leurs textes de la répression gouvernementale ; dans un nombre croissant de pays, au XIX^e siècle, la liberté d'expression était devenue une garantie constitutionnelle. Mais ces droits s'accompagnaient de devoirs. Quand des journaux ont commencé à lancer des attaques virulentes qui violaient la vie privée ou calomniaient les gens, des règles sont apparues sur la protection de la vie privée et sur la possibilité d'intenter des poursuites pour diffamation.

Ces changements dans les modes de gouvernance reflètent également une transformation plus profonde, plus fondamentale des valeurs sous-jacentes. Grâce à Gutenberg, nous avons commencé à prendre conscience du pouvoir de l'écrit – et, à terme, de l'importance de la diffusion large d'informations dans toute la société. Au cours des siècles, nous avons opté pour une augmentation des flux d'informations plutôt que le contraire, mais en nous gardant de leurs excès grâce à des règles limitant leur utilisation abusive, plus que par la censure (du moins pas en priorité).

Avec l'entrée dans l'ère des big data, la société va connaître une révolution similaire. Les big data transforment déjà de nombreux aspects de notre vie et de notre mode de pensée. Il nous faut réexaminer les principes fondamentaux qui vont favoriser leur développement tout en atténuant leurs effets néfastes potentiels. Après la révolution provoquée par l'imprimerie, nos sociétés ont eu des siècles pour procéder à des ajustements. Aujourd'hui, nous ne disposons que de peu de temps, probablement quelques années à peine...

Dans cette nouvelle ère, il ne suffira pas de procéder à de simples modifications des règles existantes pour régir les big data et en éclairer la face sombre. Plus qu'un changement de paramètres, c'est un changement de paradigme que la situation exige. La protection de la vie privée exige que les utilisateurs de big data aient à répondre plus sérieusement de leurs actes. La notion même de justice devra être redéfinie par la société de façon à garantir à chacun sa liberté d'action (et que chacun soit ainsi tenu pour responsable de ses actes). Enfin, de nouvelles institutions et de nouveaux métiers devront voir le jour, destinés à interpréter les algorithmes

complexes à la base des résultats des big data, et se mettre au service de la défense de ceux qui pourraient en être les victimes.

De la vie privée à la responsabilité

Durant des décennies, un principe essentiel a été instauré à travers le monde : créer des lois protégeant la vie privée, en accordant aux individus un droit de contrôle sur leurs informations personnelles. Autrement dit celui de décider s'ils acceptent qu'elles fassent l'objet d'un traitement, de quelle manière et par qui. À l'ère d'Internet, cet idéal louable a souvent pris la forme d'un formulaire de « notification et consentement ». À l'ère des big data, cependant, on a vu que ce mécanisme prévu pour protéger la vie privée n'était plus adapté : une bonne partie de la valeur des données se trouve en effet dans des utilisations secondaires qui n'ont pas été forcément envisagées au moment de leur collecte.

À l'ère des grands ensembles de données, nous imaginons un cadre de protection de la vie privée très différent, qui serait moins axé sur le consentement individuel au moment de la collecte que bien plus fortement fondé sur la responsabilisation des utilisateurs de données. Dans un tel cadre, les entreprises voulant réutiliser les données d'une façon particulière devront procéder à une évaluation officielle qui prenne en compte les conséquences sur les personnes du traitement des informations qui les concernent. Il ne sera pas nécessaire que cette évaluation soit détaillée au cas par cas, car la tâche serait fastidieuse ; ce que les lois futures protégeant la vie privée devront définir en effet, ce sont de larges catégories d'utilisation, notamment celles autorisées sans mesures de protection normalisées et celles avec des mesures de protection normalisées mais limitées. Pour les initiatives présentant plus de risques, les organismes de réglementation établiront des règles de base définissant la manière dont les utilisateurs de données devront évaluer les dangers de l'utilisation prévue et déterminer les meilleures mesures à prendre pour éviter ou atténuer ses effets nuisibles potentiels. Cela devrait encourager la réutilisation créative des données tout en s'assurant de mesures adéquates qui évitent de causer un préjudice aux personnes concernées.

Une utilisation de big data évaluée de façon formelle et correcte, ainsi que l'assurance de la bonne application de ses conclusions présentent des avantages tangibles pour les utilisateurs de données : ils seront alors libres

de procéder à de nombreuses utilisations secondaires de données personnelles sans avoir à revenir vers les personnes concernées pour obtenir leur consentement explicite. Par contre, la responsabilité légale des utilisateurs de données sera engagée en cas d'évaluation bâclée ou d'une mise en œuvre inadéquate des mesures de protection. Ils pourront faire l'objet d'ordonnances judiciaires, d'amendes et sans doute même de poursuites judiciaires. La responsabilité des utilisateurs de données ne peut être véritablement engagée que dans un cadre réglementaire strict.

La mise en données des « postérieurs » citée au chapitre 5 illustre bien ce qui peut se produire dans la pratique. Imaginez qu'une société ait vendu un antivol de voiture qui utilise la posture assise du conducteur comme seul identifiant. Puis qu'elle ait plus tard effectué une nouvelle analyse des mêmes informations dans le but de prédire les différents « degrés d'attention » des conducteurs en fonction de leur état – somnolent, éméché ou en colère. Et ce, afin d'envoyer un avertissement aux automobilistes des alentours de façon à empêcher des accidents. Selon les règles actuelles de la protection de la vie privée, l'entreprise n'ayant pas l'autorisation de réutiliser les informations pensera qu'il lui faut en obtenir une nouvelle auprès des conducteurs, par le biais du formulaire de notification et consentement. Mais dans le cadre d'un système fondé sur la responsabilisation des utilisateurs de données, la société n'aurait qu'à évaluer les dangers présentés par l'utilisation projetée, et si ceux-ci se révélaient minimes, elle pourrait poursuivre son projet – et améliorer la sécurité routière du même coup.

Transférer la responsabilité du public vers les utilisateurs de données est logique pour un certain nombre de raisons. Bien plus que quiconque, et certainement mieux que les consommateurs ou les organismes de réglementation, ils savent ce qu'ils comptent faire des données. En effectuant eux-mêmes l'évaluation (ou en confiant l'opération à des experts), ils éviteront le problème d'avoir à dévoiler des stratégies commerciales confidentielles à des personnes extérieures à la société. Et surtout, les utilisateurs de données récoltant la plupart des fruits des données lors de leur utilisation secondaire, il est tout à fait normal qu'ils soient tenus responsables de leurs actes et que cette tâche d'évaluation leur incombe.

Avec cet autre cadre de protection de la vie privée, les utilisateurs de données ne seront plus légalement tenus de supprimer les informations

personnelles après leur utilisation initiale, comme l'exigent la plupart des lois sur la vie privée actuellement en vigueur. Le changement est conséquent puisque, comme nous l'avons vu, seule l'exploitation de la valeur latente des données permettra de voir se multiplier les émules de Maury qui extrairont la valeur maximale pour leur bénéfice personnel – et celui de la société. Les utilisateurs de données seront dorénavant autorisés à conserver plus longtemps les informations personnelles, mais cependant pas indéfiniment. La société doit soigneusement peser le pour et le contre entre les avantages présentés par la réutilisation de données et les risques associés à la divulgation d'un trop grand nombre d'informations.

Pour trouver un juste équilibre, les organismes de réglementation pourraient définir des calendriers de réutilisation différents en fonction du risque inhérent aux données, ainsi que des valeurs différentes selon les sociétés. Certaines nations seront sans doute plus prudentes que d'autres, tout comme certains types de données pourront être considérés comme plus sensibles que d'autres. Cette méthode bannit également le spectre de la « mémoire permanente » – le risque pour une personne de ne jamais pouvoir échapper à son passé parce qu'il sera toujours possible d'exhumer les documents numériques la concernant. Sans ces mesures, nos données personnelles présentent une menace potentielle, une sorte d'épée de Damoclès qui pourrait, des années plus tard, nous transpercer sous la forme d'un détail d'ordre privé ou d'un achat regrettable. Par ailleurs, l'instauration d'un délai maximal constitue pour les utilisateurs de données une bonne incitation à s'en servir avant de les perdre. Cela établit, nous semble-t-il, un équilibre adéquat pour l'ère des big data : les sociétés obtiennent le droit d'utiliser plus longtemps les données personnelles mais en contrepartie, elles doivent en assumer la responsabilité et prendre l'engagement d'effacer ces données personnelles au bout d'un certain temps.

Outre le changement de réglementation qui fait passer la protection de la vie privée du terrain du « consentement » à celui de la « responsabilité », nous pensons qu'un autre moyen peut contribuer à protéger la vie privée dans certaines circonstances : l'innovation technique, comme cette méthode qui en est à ses débuts et dont le concept repose sur le « respect différentiel de la vie privée ». Il s'agit de brouiller délibérément les données de manière qu'une requête effectuée sur un vaste ensemble ne révèle que des résultats approximatifs et non exacts. Ainsi, tenter d'associer des points de données

spécifiques avec des personnes particulières serait une opération difficile et coûteuse.

Brouiller les informations donne l'impression que l'opération pourrait en détruire la valeur. Mais pas nécessairement ! Du moins, le compromis peut être intéressant. Par exemple, certains experts en politique technologique font remarquer que Facebook pratique une forme de respect différentiel de la vie privée lorsqu'il transmet des informations sur ses utilisateurs à des annonceurs potentiels : les nombres qu'il livre sont approximatifs, et ils ne peuvent donc pas contribuer à révéler l'identité des personnes. Une recherche sur les femmes asiatiques résidant à Atlanta qui s'intéressent à l'ashtanga yoga donnera un résultat du type « environ 400 » plutôt qu'un nombre exact, ce qui empêche d'utiliser l'information pour identifier de manière statistique quelqu'un de précis.

Contrôler l'utilisation des données en remplaçant le consentement individuel par la responsabilité de l'utilisateur constitue un changement primordial tout à fait fondamental pour régir efficacement les big data. Mais ce n'est pas le seul.

Personnes contre prévisions

Les tribunaux tiennent les personnes pour responsables de leurs actes. Quand les juges rendent des verdicts impartiaux, après un procès équitable, justice est faite. Mais à l'ère des big data, notre notion de la justice doit être redéfinie pour préserver la notion du pouvoir d'agir que possèdent les individus, c'est-à-dire l'exercice de leur libre arbitre dans le choix de leurs actions. C'est là simplement l'idée que les personnes ne peuvent et ne doivent être tenues pour responsables que de leur comportement, et non de leurs intentions.

Avant les big data, cette liberté était si évidente qu'en fait, il n'était pas vraiment nécessaire de le préciser. Après tout, c'est ainsi que notre système juridique fonctionne : nous tenons les gens pour responsables de leurs actes en examinant ceux qu'ils ont commis. Avec les big data en revanche, nous pouvons prédire les actions humaines avec de plus en plus d'exactitude. La tentation est grande alors de juger une personne non pas sur ce qu'elle a fait réellement, mais sur la base de la prévision de ce qu'elle pourrait faire.

À l'ère des big data, nous devons élargir la façon dont nous comprenons la justice, et exiger que soit incluse la sauvegarde du pouvoir

d'agir propre à l'être humain de la même manière que l'équité procédurale est actuellement préservée. Sans ces garanties, l'idée même de justice pourrait être totalement compromise.

En garantissant cette capacité d'agir, nous aurons l'assurance que le gouvernement jugera notre comportement en se fondant sur des actes réels, et pas simplement sur une analyse de données. Ainsi ne doit-il nous tenir pour responsables que de nos actes passés, et non d'actes à venir énoncés par des prévisions statistiques. Et lorsque l'État juge des actes antérieurs, il faudrait l'empêcher de se fonder uniquement sur des ensembles de données. Prenons l'exemple de neuf sociétés soupçonnées d'entente sur les prix. Il est totalement acceptable d'utiliser des analyses de big data pour déceler une éventuelle collusion de façon que les organismes de régulation puissent enquêter et monter un dossier à l'aide de méthodes classiques. Mais ces sociétés ne peuvent pas être déclarées coupables sur la simple suggestion par des big data qu'elles auraient probablement commis un délit.

En dehors du gouvernement, un principe similaire devrait être appliqué quand des entreprises prennent des décisions très importantes qui nous concernent – recrutement ou licenciement, offre d'un prêt hypothécaire ou refus d'une carte de crédit. Lorsque des décisions se fondent essentiellement sur des prévisions issues de big data, il est recommandé que certaines garanties soient mises en place. La première est la transparence : donner accès aux données et à l'algorithme sous-tendant la prévision concernant la personne. La seconde est la certification : faire certifier l'algorithme destiné à certaines utilisations sensibles par un expert indépendant qui en confirmera la fiabilité et la validité. La troisième est la réfutabilité : spécifier les moyens concrets permettant à la personne concernée par une prévision de réfuter celle-ci. (Cette démarche est analogue à la tradition qui a cours dans les sciences, celle de divulguer tout facteur susceptible de contredire les résultats d'une étude.)

Aspect essentiel, préserver le pouvoir d'action constitue une garantie contre la menace d'une dictature des données, qui consisterait à attribuer aux données une signification et une importance plus grandes qu'elles ne le méritent.

Il est tout aussi crucial pour nous de sauvegarder la responsabilité individuelle. La société sera confrontée à la grande tentation de ne plus tenir les individus pour responsables et d'opter pour une démarche de gestion des risques, autrement dit de prendre des décisions concernant les individus en

se fondant sur l'évaluation des possibilités et des probabilités de résultats potentiels. Avec une telle quantité de données disponibles, apparemment objectives, il peut paraître séduisant de soustraire la prise de décision à toute influence des émotions et de la subjectivité. Il s'agirait alors de s'appuyer sur des algorithmes plutôt que sur les appréciations subjectives de juges et d'évaluateurs pour que les décisions ne soient pas prises en termes de responsabilité personnelle mais en fonction de risques plus « objectifs » à prévenir.

À titre d'exemple, revenons sur ce que nous avons déjà exposé, à savoir que les big data sont une véritable invitation à prédire qui est susceptible de commettre un crime et à soumettre constamment ces personnes à une surveillance particulière au nom de la nécessité de réduire les risques. Ceux qui se retrouveraient classés dans cette catégorie pourraient avoir le sentiment, à juste titre, qu'ils sont punis arbitrairement sans aucune enquête préalable alors qu'ils ne peuvent pas être tenus responsables d'un acte réel. Imaginez qu'un algorithme détecte chez un adolescent particulier une très forte probabilité de commettre un crime dans les trois années à venir. Les autorités désignent alors un travailleur social pour lui rendre visite une fois par mois, afin de « garder un œil sur lui » et d'essayer de l'aider à éviter toute infraction.

Si l'adolescent, les membres de sa famille, ses amis, ses professeurs ou ses employeurs ressentent les visites comme une stigmatisation, ce qui n'est pas impossible, l'intervention elle-même prend alors l'allure d'une punition, d'une sanction pour un délit qui n'a pourtant pas été commis ! Et la situation n'est pas meilleure non plus quand les visites ne sont pas perçues comme une punition, mais simplement comme une tentative de réduire la probabilité de problèmes futurs – comme un moyen de minimiser le risque (en l'occurrence ici, le risque d'un crime qui compromettrait la sécurité publique) ! Plus nous nous appuyons, pour réduire les risques au sein de la société, sur des interventions fondées sur les données plutôt que sur la responsabilité réelle de quelqu'un, plus nous dévaluons l'idéal de la responsabilité individuelle. L'État prédictif, c'est bien plus encore que l'État providence. Priver quelqu'un de la responsabilité de ses actes, c'est lui ôter sa liberté fondamentale, celle de choisir ses comportements.

Si l'État fonde de nombreuses décisions sur des prévisions et sur la volonté de limiter les risques, nos choix individuels – et donc notre liberté d'action individuelle – n'ont plus aucune importance. Sans la notion de

culpabilité, il ne peut y avoir de notion d'innocence. Adhérer à cette démarche n'améliorerait pas notre société, au contraire, elle l'appauvrirait.

Un pilier fondamental du contrôle des big data doit être la garantie que nous continuerons à prendre en compte la responsabilité personnelle et le comportement réel des individus avant de les juger, et non en nous fondant sur une analyse « objective » des données qui seule déterminerait si ce sont des coupables potentiels. C'est seulement de cette manière que nous traiterons les individus comme les êtres humains qu'ils sont, c'est-à-dire des personnes libres de choisir leurs actes et qui ont le droit de n'être jugées que sur ces actes.

Briser la boîte noire

Les systèmes informatiques fondent actuellement leurs décisions sur des règles qu'ils ont été explicitement programmés pour suivre. Ainsi, lorsqu'un problème se produit à la suite d'une décision prise par l'ordinateur, et cela arrive inévitablement de temps à autre, il est possible de revenir en arrière et de comprendre pourquoi il l'a prise. Par exemple, nous pouvons étudier pourquoi le système de pilotage automatique a fait cabrer l'avion de cinq degrés quand un capteur externe a détecté une brusque augmentation du taux d'humidité. Un code informatique peut aujourd'hui être ouvert et inspecté, et ceux qui savent l'interpréter peuvent retrouver et comprendre l'origine des décisions de l'ordinateur, qu'elle qu'en soit la complexité.

Mais avec l'analyse des big data, cette traçabilité deviendra beaucoup plus difficile. La base de prévisions fournies par un algorithme sera souvent bien trop complexe pour être comprise par la plupart des gens.

Quand les ordinateurs étaient explicitement programmés pour suivre des ensembles d'instructions, comme avec le programme de traduction du russe vers l'anglais mis au point par IBM dès 1954, une personne pouvait facilement comprendre pourquoi le logiciel avait remplacé un mot par un autre. Mais Google Traduction intègre des milliards de pages de traductions pour décider du choix d'un terme ou un autre, par exemple le mot anglais *light* peut être traduit en français par le terme « lumière » ou l'adjectif « léger » (il faut donc savoir si l'on fait ici référence à la clarté ou au poids). Il est impossible à qui que ce soit d'expliquer les raisons précises du choix

des termes par le programme parce qu'ils sont fondés sur des quantités énormes de données et de lourds calculs statistiques.

Les big data fonctionnent à une échelle qui dépasse notre compréhension ordinaire. Par exemple, la corrélation identifiée par Google entre une poignée de requêtes et la grippe fut le résultat de tests effectués sur des mots à l'aide de 450 millions de modèles mathématiques. Par contre, Cynthia Rudin avait conçu à l'origine 106 variables prédictrices d'un incendie dans un trou d'homme, et elle fut en mesure d'expliquer aux dirigeants de Con Edison pourquoi son programme indiquait un ordre de priorité pour l'inspection des sites. « L'explicabilité », comme on l'appelle dans les milieux de l'intelligence artificielle, est importante pour nous, simples mortels, qui avons tendance à vouloir connaître le pourquoi, et pas simplement le quoi. Mais que se passerait-il si, au lieu de 106 variables prédictrices, le système en générerait automatiquement 601, si la grande majorité de ce nombre astronomique avait une faible pondération et si, réunies, elles amélioreraient l'exactitude du modèle ? La base de toute prévision pourrait être incroyablement complexe. Que pourrait-elle dire alors aux dirigeants pour les convaincre de réaffecter leurs budgets limités ?

Ces scénarios nous permettent de voir le risque que les prévisions des big data, ainsi que les algorithmes et les ensembles de données qui les sous-tendent, puissent se transformer en boîtes noires dont nous ne pourrions attendre aucune explication, ni aucune traçabilité et en lesquelles nous ne pourrions avoir aucune confiance. Pour éviter cela, les big data exigeront contrôle et transparence, ce qui demandera de nouveaux types d'expertise et d'institutions. Ces nouveaux acteurs apporteront leur contribution dans des domaines où la société a besoin d'examiner de près les prévisions des big data et de donner aux personnes qui s'estiment lésées par elles le moyen de demander réparation.

En tant que société, nous avons souvent vu apparaître de nouvelles entités de ce type avec un besoin urgent d'experts capables de gérer de nouvelles techniques parce que la complexité et la spécialisation ont augmenté de façon spectaculaire dans un domaine particulier. Des professions liées à des domaines comme le droit, la médecine, la comptabilité et l'ingénierie ont subi cette mutation il y a plus d'un siècle. Plus récemment, des spécialistes de la sécurité informatique et de la protection de la vie privée sont apparus pour certifier que les entreprises se conforment aux bonnes pratiques définies par des organismes comme

l'Organisation internationale de normalisation (elle-même créée pour répondre à un nouveau besoin de directives dans ce domaine).

Les big data nécessiteront un nouveau type d'experts pour assumer ce rôle. Sans doute les appellera-t-on « algorithmistes ». Ils pourraient se présenter sous deux formes – des entités indépendantes pour contrôler les entreprises de l'extérieur, et des employés ou des services intégrés pour les contrôler de l'intérieur – tout comme les sociétés ont des comptables en interne et des auditeurs externes pour examiner les comptes.

La naissance de l'algorithmiste

Ces nouveaux professionnels seraient des experts dans les domaines de la science informatique, des mathématiques et des statistiques ; leur rôle consisterait à examiner les analyses et les prévisions de big data. Les algorithmistes prêteraient serment d'impartialité et de respect du secret professionnel, comme le font actuellement les comptables et certains autres professionnels. Ils évalueraient le choix des sources de données, celui des outils servant à l'analyse et aux prévisions, notamment les algorithmes et les modèles, et l'interprétation des résultats. En cas de litige, ils auraient accès aux algorithmes, aux méthodes statistiques et aux ensembles de données à l'origine d'une décision donnée.

S'il y avait eu un algorithmiste dans le personnel du Département de la sécurité intérieure des États-Unis en 2004, il aurait pu empêcher l'agence de publier une liste de passagers interdits de vol tellement erronée que le sénateur Kennedy y figurait. On peut trouver d'autres exemples plus récents où des algorithmistes auraient pu jouer un rôle au Japon, en France, en Allemagne et en Italie, où des personnes se sont plaintes que la fonction de « saisie semi-automatique » de Google, qui suggère une liste de mots-clés associés au nom saisi, les avait diffamées ! La liste est essentiellement fondée sur la fréquence de recherches antérieures : les termes sont classés selon leur probabilité mathématique. Qui d'entre nous ne serait pas furieux si le mot « détenu » ou « prostituée » s'affichait à côté de son nom quand un partenaire en affaires ou un amoureux potentiel fait une recherche sur Internet pour prendre des renseignements sur nous ?

Tels que nous les imaginons, les algorithmistes correspondraient à une réponse du marché pour aborder les problèmes évoqués plus haut, et cela pourrait contribuer à éviter le recours à des formes plus intrusives de

réglementation. Quand un déluge d'informations financières a dû être géré au début du xx^e siècle, sont apparus les comptables et les auditeurs. Les algorithmistes répondraient aujourd'hui à un besoin similaire. Il n'a pas été facile de comprendre que des spécialistes à l'organisation flexible et autorégulée étaient nécessaires à la gestion d'une énorme masse de chiffres. Le marché a réagi en donnant naissance à un nouveau secteur d'entreprises qui sont entrées en compétition sur le nouveau créneau de la surveillance financière. En proposant ce service, ce nouveau type de professionnels a renforcé la confiance de la société dans l'économie. Ainsi, les big data pourraient et devraient bénéficier d'un regain de confiance similaire grâce au travail des algorithmistes.

Algorithmistes externes

Nous imaginerions bien des algorithmistes externes agir en tant que vérificateurs impartiaux des comptes pour contrôler l'exactitude ou la validité des prévisions de big data à la demande du gouvernement, par ordonnance du tribunal ou en vertu d'un règlement. Ils pourraient également recruter comme clients des sociétés de big data et réaliser des audits pour des entreprises qui ont besoin de faire appel à des experts. Ils pourraient certifier la validité des applications de big data comme les techniques antifraude ou les systèmes de placement en Bourse. Enfin, les algorithmistes externes seraient en mesure de conseiller les agences gouvernementales sur les meilleures utilisations de big data dans le secteur public.

Comme cela est le cas dans les domaines professionnels de la médecine, du droit et autres métiers, nous imaginons que cette nouvelle profession serait régie par un code de déontologie. L'impartialité, l'obligation de confidentialité, la compétence et le professionnalisme des algorithmistes seraient des conditions obligatoires régies par des règles strictes en matière de responsabilité ; en cas de non-respect de ces normes, les algorithmistes s'exposeraient à des poursuites. Ils pourraient également être appelés à témoigner en qualité d'experts lors de procès ou bien être nommés en tant qu'experts judiciaires par les juges pour les conseiller sur des aspects techniques dans des affaires particulièrement complexes.

Par ailleurs, les personnes qui estiment avoir été lésées par des prévisions de big data – un patient qu'on refuse d'opérer, un détenu qui

s'est vu refuser sa mise en liberté conditionnelle, le requérant à qui on refuse une hypothèque – pourraient se tourner vers les algorithmistes comme elles se tournent déjà vers des avocats pour qu'ils les aident à comprendre ces décisions et à interjeter appel.

Algorithmistes en interne

Les algorithmistes en interne travailleraient au sein d'une entreprise pour y surveiller les activités en matière de big data. Ils défendraient non seulement les intérêts de la société, mais également ceux des personnes lésées par les analyses de big data réalisées dans cette société. Ils surveilleraient les activités de big data, et ils seraient l'interlocuteur privilégié de toutes ces personnes. Ils examineraient également les analyses de big data pour en vérifier l'intégrité et l'exactitude avant qu'elles ne soient mises en fonctionnement. Pour jouer le premier de ces deux rôles, les algorithmistes devront avoir un certain niveau de liberté et d'impartialité au sein de l'entreprise pour laquelle ils travaillent.

L'idée qu'une personne puisse rester impartiale au sujet des activités de la société pour laquelle elle travaille peut paraître paradoxale, mais ces situations sont en fait relativement courantes. Les conseils de surveillance dans les grandes institutions financières en sont un exemple ; ainsi les conseils d'administration de nombreuses entreprises sont responsables vis-à-vis des actionnaires et non vis-à-vis de la direction. De nombreuses entreprises de médias, notamment le *New York Times* et le *Washington Post*, emploient des médiateurs dont la première responsabilité est de sauvegarder la confiance du public. Ces employés traitent les réclamations des lecteurs et rappellent souvent à l'ordre leur employeur publiquement quand ils établissent que celui-ci a causé du tort.

Il existe actuellement une profession assez proche de celle de l'algorithmiste interne – celle d'un expert chargé de veiller à ce que les informations personnelles ne soient pas utilisées à mauvais escient dans le cadre d'une entreprise. Par exemple, l'Allemagne impose aux sociétés au-delà d'une certaine taille, c'est-à-dire comptant généralement dix personnes ou plus affectées au traitement des informations personnelles, de désigner un représentant pour la protection des données. Depuis les années 1970, ces représentants internes ont développé leur éthique professionnelle et un esprit de corps. Ils se réunissent pour partager leurs expériences en matière

de bonnes pratiques et de formations ; et ils disposent de leurs propres médias et organisent leurs propres conférences professionnelles. De plus, ils ont réussi à maintenir une double allégeance, à la fois à leur employeur et à leurs fonctions de contrôleurs impartiaux ; ils sont capables d’agir comme médiateurs en matière de protection des données tout en intégrant dans l’ensemble des opérations de leur entreprise les valeurs de respect de la vie privée et de la protection des informations personnelles. Nous sommes convaincus que les algorithmistes internes pourraient parfaitement jouer ce même rôle.

Contrôler les « barons » des données

Les données sont à la société de l’information ce que le carburant fut à l’économie industrielle : une source essentielle d’inspiration pour l’innovation. La créativité et la productivité potentielles seraient probablement restreintes sans l’existence de données riches et dynamiques et d’un marché des services robuste.

Dans ce chapitre, nous avons exposé trois nouvelles stratégies fondamentales de gouvernance des big data, portant sur la protection de la vie privée, les intentions et la vérification des algorithmes. Nous sommes convaincus que, si elles étaient mises en œuvre, cela limiterait la face sombre des big data. Mais lorsque le secteur naissant des big data se développera, un autre défi de taille sera à relever, celui de garantir la compétitivité des marchés. Au XXI^e siècle, nous devons empêcher l’émergence de barons des données, version moderne des requins du XIX^e siècle qui avaient dominé les réseaux du rail, de la fabrication de l’acier et du télégraphe aux États-Unis.

Pour contrôler ces industriels des premiers temps, les États-Unis avaient mis en place des règles antitrust extrêmement adaptables. Conçues à l’origine pour les chemins de fer dans les années 1800, elles ont été plus tard appliquées à des sociétés qui étaient les « gardiennes » des flux d’informations dont dépendaient les entreprises : National Cash Register dans les années 1910, IBM dans les années 1960 et, plus tard, Xerox dans les années 1970, AT&T dans les années 1980, Microsoft dans les années 1990 et aujourd’hui, Google. Les technologies que ces sociétés ont lancées sont devenues les composantes essentielles de l’« infrastructure de

l'information » dans l'économie, et il a fallu déterminer des contraintes légales pour contrer le risque de domination malsaine du marché.

Pour garantir des conditions propices à un marché florissant des big data, il nous faudra prendre des mesures comparables à celles mises en place pour régir la concurrence et la surveillance lors de l'utilisation de technologies antérieures. Nous devons favoriser les transactions de données, en développant les concessions de licences d'exploitation et en assurant l'interopérabilité des données. Cela soulève la question de savoir si la société aurait avantage à instituer un « droit d'exclusion » portant sur les données, qui soit bien pensé et soigneusement formulé (similaire aux droits de propriété intellectuelle, aussi provocante que paraisse cette suggestion !). Assurément, cette entreprise constituerait un défi de taille pour les décideurs politiques et comporterait en outre des risques importants pour les autres.

De toute évidence, il est impossible de prédire la manière dont évoluera une technologie ; les big data elles-mêmes ne sauraient prédire leur propre évolution ! Les organismes de réglementation devront trouver dans leurs interventions un juste équilibre entre la prudence et l'audace – et l'histoire des lois antitrust suggère une voie possible pour atteindre cet objectif.

La législation antitrust a réprimé les abus de pouvoir. Mais, chose remarquable, ses principes ont pu merveilleusement s'adapter d'un secteur à l'autre, et dans différents types d'industries de réseaux. C'est exactement cette forme de réglementation vigoureuse – qui ne privilégie pas un type de technologie plutôt qu'un autre – qui est utile, étant donné qu'elle protège la concurrence, sans aller au-delà. Par conséquent, cette législation antitrust peut aider le secteur des big data à aller de l'avant tout comme elle l'a fait pour les chemins de fer. Les gouvernements, qui sont parmi les plus gros détenteurs de données du monde, devraient également rendre leurs données accessibles au public. Signe encourageant, certains le font déjà – dans une certaine mesure, du moins.

La leçon à tirer de la législation antitrust est qu'une fois que les principes de base ont été définis, les organismes de réglementation peuvent les appliquer pour garantir les mesures de protection et le soutien adéquats. De même, les trois stratégies que nous avons exposées – la responsabilité des utilisateurs de données plutôt que le consentement individuel pour la protection de la vie privée ; la sauvegarde du pouvoir d'agir dans un contexte de prévisions ; et la création d'une nouvelle caste professionnelle

de vérificateurs de big data que nous avons appelés algorithmistes – pourraient servir de base pour une gouvernance efficace et équitable de l'information à l'ère des big data.

Dans de nombreux domaines, depuis la technologie nucléaire jusqu'à la biotechnologie, nous avons commencé par inventer des outils et ce n'est que plus tard, en découvrant leur nocivité potentielle, que nous avons entrepris de nous protéger et de concevoir des mécanismes de sécurité nécessaires. À cet égard, les big data viennent s'ajouter à d'autres défis sans solutions idéales, et nous amènent à nous interroger à nouveau sur la manière que nous avons de régir les divers aspects du monde. À chaque génération d'aborder à son tour ces questions. Notre tâche est d'évaluer les risques de cette technologie puissante, d'aider à son développement – et d'en tirer profit. Tout comme la presse à imprimer a en son temps bouleversé la façon dont la société se régissait, les big data exerceront elles aussi une influence du même ordre sur la société. Cela nous obligera à relever de nouveaux défis en apportant de nouvelles réponses. Pour assurer la protection des individus tout en faisant la promotion de cette technologie des big data, nous ne devons pas laisser son développement échapper au contrôle humain !

10

Demain

Au début des années 2000, Mike Flowers est avocat au bureau du procureur du district de Manhattan ; sa tâche consiste à engager des poursuites de toutes sortes, depuis les homicides jusqu'aux crimes financiers à Wall Street ; plus tard, il intègre un riche cabinet d'avocats d'affaires. Mais comme il s'y ennue également, au bout d'un an, il décide de se trouver une activité plus motivante ; il en vient à penser qu'il pourrait lui être possible de participer au travail de reconstruction de l'Irak. Grâce aux recommandations de l'un de ses collègues auprès de personnes influentes, Flowers fait quelque temps après route pour la Zone verte, enclave hautement sécurisée dans le centre de Bagdad où étaient à l'époque cantonnées les troupes américaines, et intègre l'équipe juridique dans le cadre du procès de Saddam Hussein.

La plus grande partie de sa mission se révèle alors être d'ordre logistique, et non juridique. Flowers doit identifier les zones susceptibles de receler des fosses communes afin d'y envoyer des enquêteurs faire des fouilles. Il doit aussi trouver les bons itinéraires pour faire passer des témoins vers la Zone verte sans qu'ils se fassent souffler par quelque engin explosif improvisé, la dure réalité quotidienne des attaques à Bagdad. Il fait le constat que l'armée traite ces tâches comme des problèmes d'information et que les données leur sont d'un grand secours ; les analystes du renseignement combinent des rapports de terrain avec des détails sur les attaques précédentes – lieu, heure et nombre de blessés victimes de ces engins – pour prédire le chemin le plus sûr à prendre ce jour-là.

À son retour à New York quelques années plus tard, Flowers comprend que ces analyses pourraient représenter, pour combattre le crime, un moyen plus efficace que ceux qu'il a eus à sa disposition en tant que procureur. Et

il rencontre quelqu'un qui partage complètement son avis en la personne du maire de la ville, Michael Bloomberg, qui a fait fortune en fournissant des informations financières aux banques. Flowers est alors nommé dans une unité spéciale chargée d'analyser des chiffres susceptibles de démasquer les escrocs responsables du scandale des « subprimes^a » en 2009. L'unité remporte de tels succès qu'un an plus tard, le même maire lui demande d'élargir son champ d'action. Flowers devient ainsi le premier « directeur du service Analyse » de la ville ; il a pour mission de constituer une équipe composée des meilleurs scientifiques spécialistes des données afin que les mines d'informations stockées par la ville dans tous les domaines, mais jusque-là inexploitées, soient utilisées avec une efficacité maximale.

Flowers ratisse large pour trouver les bons candidats. « Je ne m'intéressais pas du tout aux statisticiens chevronnés, déclare-t-il aujourd'hui. Je craignais qu'ils ne soient réticents à adopter cette nouvelle méthode de résolution de problèmes. » Lors de ses entretiens avec des statisticiens classiques dans le cadre du projet relatif aux fraudes fiscales, ceux-ci ont effectivement eu tendance à exprimer leurs préoccupations concernant les méthodes mathématiques. « Je n'avais même pas réfléchi au modèle que j'allais utiliser. Je voulais une idée exploitable, c'était tout ce qui m'intéressait », dit-il. Finalement, il sélectionne une équipe de cinq personnes qu'il rebaptise les « gosses ». Tous, à l'exception d'un seul, ont fait des études d'économie, qu'ils ont achevées depuis tout juste un an ou deux ; ils n'ont pas encore l'expérience des grandes villes mais on sent chez tous un réel potentiel de créativité.

L'un des premiers problèmes auxquels l'équipe se retrouve confrontée est celui des « conversions illégales » de logements – pratique consistant à subdiviser une habitation en plusieurs petites unités de façon à y loger jusqu'à dix fois plus de personnes que prévu, ce qui entraîne des risques majeurs d'incendie, transforme ces lieux en de véritables foyers du crime, de la drogue sans oublier la prolifération d'animaux nuisibles. Il faut se représenter les lieux avec des enchevêtrements de rallonges électriques qui serpentent le long des murs, des plaques chauffantes posées en équilibre instable sur des couvre-lits et les gens vivant là si entassés que les incendies y font régulièrement de nombreuses victimes ; en 2005, deux pompiers ont péri en tentant de secourir des résidents. La ville de New York reçoit chaque année à peu près 25 000 plaintes pour conversion illégale de logement, mais elle ne dispose que de 200 inspecteurs pour les traiter. À première vue, il

n'y a aucun moyen pratique de faire la distinction entre les cas banals d'incendie et ceux où il risque vraiment de se déclencher, vu l'état des lieux. Aux yeux de Flowers et de sa bande de jeunes, le problème devait pouvoir être résolu grâce à de bonnes quantités de données.

Ils ont commencé par établir la liste de tous les bâtiments de la ville – 900 000 en tout. Ensuite, ils ont introduit dans le système des ensembles de données provenant de 19 administrations différentes et indiquant par exemple si le propriétaire du bâtiment était en retard de paiement de taxes foncières, s'il y avait eu des procédures de saisie et si des anomalies dans la consommation des produits de services publics ou bien des défauts de paiement avaient entraîné des coupures de ces services. Ils y ont ajouté des informations sur le type de bâtiment, la date de construction, ou encore le nombre d'ambulances qui ont eu à intervenir à cette adresse, le taux de criminalité du lieu, les plaintes liées à la présence de rongeurs, etc. Puis ils ont comparé toutes ces informations avec les données relatives aux incendies sur une période de cinq ans, classées selon la gravité du sinistre, et recherché des corrélations de manière à produire un système capable de prédire quelles plaintes devaient faire l'objet d'une enquête en urgence.

À l'origine, une bonne partie des données ne se sont pas présentées pas sous une forme utilisable. Par exemple, il n'y a pas eu au cours des décennies une seule méthode, normalisée, pour décrire un lieu dans les dossiers concernant la ville ; chaque administration et chaque service ont employé la leur. La direction des bâtiments a attribué à chaque structure un numéro de bâtiment unique. Le service de la conservation foncière a utilisé un système de numérotation différent. Le service fiscal a donné à chaque bien immobilier un identifiant fondé sur l'arrondissement, le pâte de maisons et le terrain. La police a utilisé les coordonnées cartésiennes. Les sapeurs-pompiers s'appuient sur un système de proximité avec des « bornes d'appel d'urgence » reliées à leurs casernes, même si une borne est hors service. Les « gosses » de Flowers ont pris en compte ce désordre et conçu un système qui commence par identifier les bâtiments en utilisant une petite zone placée à l'avant de la propriété qui tient compte des coordonnées cartésiennes. Puis le système a extrait les données des autres administrations en les géolocalisant. Une méthode inexacte en soi, mais la vaste quantité de données à disposition a largement compensé les imperfections.

L'équipe ne s'est pas contentée d'effectuer des calculs. Elle s'est ensuite rendue sur le terrain pour observer le travail des inspecteurs. L'idée étant de poser aux experts toutes sortes de questions et de prendre beaucoup de notes. Quand un inspecteur en chef grisonnant s'est avisé de marmonner que le bâtiment qu'ils allaient inspecter ne poserait pas de problème, les « geeks » ont voulu savoir pourquoi il en était si sûr. Lui-même ne savait pas au juste, mais les gosses ont fini par comprendre que son intuition était fondée sur l'observation de travaux de maçonnerie récents visibles sur la façade du bâtiment, interprétés comme un signe montrant que le propriétaire se souciait d'entretenir le lieu.

De retour à ses bureaux-cabines, l'équipe s'est interrogée pour savoir comment intégrer « travaux de maçonnerie récents » dans le modèle en tant que signal. Les briques n'étant pas – du moins pas encore – mises en données. Il faut bien sûr une autorisation de la ville pour effectuer des travaux de maçonnerie en façade. L'équipe a donc ajouté l'information figurant sur l'autorisation. Ce qui a permis d'améliorer la performance prédictive du système et d'indiquer que certaines propriétés suspectées ne présentaient probablement *pas* de risque majeur.

L'analyse a fait la démonstration que des centaines de méthodes anciennes n'étaient pas les meilleures, tout comme les dénicheurs de talents dans le film *Le Stratège* ont dû admettre que leur intuition montrait quelque faiblesse. Par exemple, le nombre d'appels pour plainte reçus au numéro 311, la ligne gratuite d'urgence de la ville, a été interprété comme une indication sur la nécessité d'accorder une attention particulière à certains bâtiments. Lorsque le nombre d'appels augmente, cela signifie que la situation s'aggrave. Mais attention, cette mesure s'est révélée trompeuse. En effet, si la vue d'un rat dans le quartier chic de l'Upper East Side peut déclencher trente appels en une heure, il faudrait sans doute l'assaut d'un bataillon de rongeurs avant que les résidents du Bronx ne ressentent le besoin de composer le 311 ! De même, la majorité des plaintes concernant un logement illégalement converti peut avoir trait au bruit, et pas forcément traduire un danger véritable.

En juin 2011, Flowers et ses gosses ont mis leur système en route. Chaque plainte entrant dans la catégorie des conversions illégales de logement a été traitée à un rythme hebdomadaire. Ils ont rassemblé celles allant dans les 5 % en haut de la liste et concernant les risques d'incendie, et

les ont transmises aux inspecteurs pour un suivi immédiat. Stupéfaction, au vu des résultats.

Avant de recourir à l'analyse de big data, les inspecteurs traitaient les plaintes qu'ils jugeaient les plus sérieuses, mais seuls 13 % des cas correspondaient à des conditions suffisamment graves pour qu'ils émettent un ordre d'expulsion. Avec l'analyse, le taux a bondi à 70 % des bâtiments inspectés ; l'efficacité des inspecteurs a été multipliée par cinq, car les big data leur indiquent les bâtiments qui nécessitent le plus leur attention. Leur travail est devenu plus gratifiant : ils se concentrent sur les problèmes les plus importants. Cette augmentation de l'efficacité des inspecteurs a eu d'autres retombées. Tout particulièrement chez les pompiers, car dans le contexte de conversions illégales, les incendies risquent de faire quinze fois plus de victimes (morts et blessés) ! Flowers et son équipe ont eu l'air de magiciens dont la boule de cristal permet de lire l'avenir et de prédire quels lieux sont les plus à risque. Rappelons que leur démarche a consisté à prendre des quantités massives de données qui traînaient depuis des années, et qui n'avaient pratiquement pas servi depuis leur collecte, puis qu'ils les ont exploitées d'une manière nouvelle pour en extraire une réelle valeur. L'utilisation d'un volumineux corpus d'informations leur a permis de détecter des connexions qu'on ne pouvait pas repérer dans des quantités plus petites – et cela, c'est l'essence même des big data.

L'expérience des « alchimistes de l'analyse » dans la ville de New York souligne plusieurs des thèmes évoqués dans ce livre. C'est une quantité monumentale de données qu'ils ont utilisée, et pas seulement une poignée ! En fait, la liste des bâtiments de la ville qu'ils ont dressée représente ni plus ni moins que $N = \text{tous}$. Ils ont dû faire face à des données certes déstructurées, comme par exemple les informations relatives à la localisation ou aux dossiers d'ambulances, mais cela ne les a pas découragés. En fait, utiliser plus de données a eu l'avantage inhérent de compenser l'inconvénient de ne pas disposer d'informations impeccables. C'est l'énorme quantité de caractéristiques de la ville mises en données (même de façon disparate) qui a permis la réussite de l'entreprise, c'est cela qui leur a permis de traiter les informations.

L'expert et son utilisation des indices ont dû céder le pas à la méthode axée sur les données. En même temps, Flowers et ses gosses n'ont cessé de tester leur système auprès d'inspecteurs chevronnés, et l'ont amélioré en puisant dans leur expérience. Mais le succès du programme tient à une autre

raison, plus importante : il repose sur la corrélation et la causalité en a été écartée.

« Je ne m'intéresse pas à la causalité sauf si son importance est avérée, explique Flowers. La causalité, c'est pour les autres, et franchement, il est hasardeux de se mettre à parler de causalité. Je ne pense pas qu'il y ait le moindre lien de cause à effet par exemple entre une procédure de saisie de biens engagée un jour par quelqu'un et le fait que le lieu concerné présente ou pas (à cause d'antécédents) un risque marquant d'incendie pour le bâtiment. Selon moi, il serait stupide de raisonner ainsi et personne ne pourrait réellement dire cela. On penserait plutôt que ce sont des facteurs sous-jacents. Mais je ne veux même pas entrer dans ces considérations. Ce qu'il me faut, c'est un point de données spécifique dont je puisse disposer et qui ait du sens. C'est alors qu'il sera possible d'en tenir compte pour agir. Sinon, nous ne le ferons pas. Vous savez, nous avons de vrais problèmes à résoudre. Franchement, en ce moment, je n'ai pas de temps à perdre en pensant à d'autres choses comme la causalité. »

Quand les données parlent

Sur un plan pratique, les effets des big data sont importants, du fait que cette technologie doit résoudre des problèmes quotidiens complexes. Mais ce n'est que le début. Les big data vont transformer notre mode de vie, notre façon de travailler et de penser. Les changements à venir sont d'une certaine manière encore plus grands que ceux dus aux innovations antérieures, qui avaient déjà considérablement accru l'importance des informations (leur portée, leur ampleur...) pour la société. Le sol remue sous nos pieds ! Exit les anciennes certitudes. À cause des big data, il faut ouvrir de nouveaux débats sur la nature de la prise de décision, du destin et de la justice. Est remise en question la vision du monde en termes de causes et c'est à une prépondérance de corrélations que nous avons désormais affaire. Avoir la connaissance signifiait jusqu'à présent qu'on comprenait le passé, cela va bientôt vouloir dire qu'on est capable de prévoir l'avenir.

Ces questions sont bien plus importantes que celles de l'époque où nous nous apprêtions à exploiter le filon du commerce électronique, à introduire Internet dans notre vie, à entrer dans l'ère de l'informatique ou à nous mettre à utiliser des abaquas. Elles sont fondamentales pour notre société et notre existence : à preuve, cette idée que notre quête de la compréhension

des causes pourrait être exagérée et qu'il y aurait souvent avantage à y renoncer (éviter le *pourquoi* et lui préférer le *quoi*). Il n'y a sans doute pas de réponse toute faite aux défis que posent les big data mais ils font partie du débat éternel sur la place de l'homme dans l'univers et sur sa quête de sens au milieu du chaos d'un monde incompréhensible.

Les big data marquent le moment où la « société de l'information » respecte enfin la promesse qu'implique son nom. Les données occupent une place de premier plan. Tous ces bits numériques que nous avons rassemblés peuvent maintenant être exploités de façon innovante pour atteindre des objectifs nouveaux et créer de nouvelles formes de valeur. Mais cela demande une nouvelle façon de penser et met en cause nos institutions et même notre identité. Une seule certitude, le volume de données ne va cesser de croître, de même que la puissance nécessaire au traitement de toutes ces données. Mais ce dont nous sommes convaincus, c'est que l'accent mis jusqu'à présent sur la technologie, le matériel ou les logiciels, doit aller ailleurs : sur ce qui se produit lorsque les données parlent.

Nous sommes en mesure de recueillir et d'analyser plus d'informations que jamais auparavant. Ce n'est plus la rareté des données qui caractérise nos tentatives d'interprétation du monde. Nous pouvons en utiliser immensément plus et, dans certains cas, presque atteindre le tout. Il nous faut alors opérer au moyen de méthodes originales et, en particulier, modifier la conception que nous avons de ce qui constitue une information utile.

Au lieu de vouloir à tout prix obtenir des données précises, exactes, épurées et rigoureuses, nous pouvons nous permettre un certain jeu. Il ne s'agit pas d'accepter des données carrément erronées ou fausses, mais de tolérer un certain désordre si cela autorise à recueillir un ensemble de données plus complet. En fait, dans certains cas, il peut même y avoir avantage à ce qu'il y ait du volume et du désordre : en effet, lorsque nous avons utilisé une seule petite portion de données exactes, ce qui nous a manqué au bout du compte, ce sont tous ces détails qui abondent en informations.

Parce qu'il est bien plus rapide et bien moins cher de trouver des corrélations que des relations de cause à effet, il est souvent préférable d'opter pour les premières. Nous continuerons à avoir besoin dans certains cas de faire appel à des études causales et à des expériences contrôlées à l'aide de données soigneusement organisées, comme par exemple lorsqu'on

veut concevoir une pièce d'avion essentielle. Mais pour bon nombre de nos besoins quotidiens, connaître le *quoi* et non le *pourquoi* est bien suffisant. Et les corrélations issues de big data peuvent indiquer la voie vers des domaines prometteurs où l'on pourra ensuite explorer des relations de causalité.

Ces corrélations, grâce à la rapidité qui les caractérise, contribuent à nous faire économiser de l'argent sur des billets d'avion, nous permettent de prédire des épidémies de grippe et de détecter les trous d'homme ou les bâtiments surpeuplés qu'il faut inspecter en priorité dans des conditions d'effectif insuffisant. Elles permettent aux compagnies d'assurance maladie de proposer une couverture sans examen médical préalable et aux prestataires de santé de cibler les patients à qui rappeler à moindre coût qu'il leur faut prendre leur traitement. On traduit des langues étrangères et l'on fait fonctionner des voitures sans pilote sur la base des corrélations issues des big data et des prédictions qu'elles donnent. Grâce à cela, Walmart peut apprendre quelle est la saveur à privilégier pour les biscuits Pop-Tarts qui seront exposés à l'entrée de ses magasins avant un ouragan – fraise, en l'occurrence ! Naturellement, il n'est pas mauvais de connaître le lien de causalité, lorsqu'on peut le trouver ! Le problème est que cela se révèle souvent difficile, et quand nous pensons l'avoir trouvé, ce n'est la plupart du temps qu'une pure illusion.

De nouveaux outils, depuis les processeurs plus rapides et une mémoire plus grande jusqu'aux logiciels et aux algorithmes ingénieux, ne suffisent pas à eux seuls à expliquer pourquoi nous sommes capables de réaliser toutes ces performances. Même si les outils ont leur importance, la raison plus fondamentale est que nous disposons de données plus nombreuses, du fait qu'un nombre croissant d'aspects du monde sont mis en données. Certes, l'ambition humaine de quantifier le monde remonte à bien avant la révolution informatique. Mais les outils numériques facilitent énormément cette mise en données. Non seulement les téléphones mobiles peuvent connaître l'identité des destinataires de nos appels et nos déplacements, mais les données qu'ils collectent peuvent servir à détecter le moment où on tombe malade. Le jour n'est pas loin où ils pourront dire si nous sommes amoureux !

Notre aptitude à innover, à accomplir plus, mieux et plus rapidement a le pouvoir de mettre au jour une valeur immense, de faire apparaître de nouveaux gagnants et de nouveaux perdants. Une grande partie de la valeur

des données proviendra de leurs utilisations secondaires, c'est-à-dire leur valeur d'option, et non pas uniquement de leurs utilisations initiales comme nous tendons à le penser habituellement. Il semble donc judicieux, pour la plupart des types de données, d'en collecter le plus possible et de les conserver tant qu'elles offrent de la valeur ajoutée. Il faudra aussi confier leur analyse à d'autres si ceux-ci sont plus aptes à en extraire la valeur (et à condition de partager les bénéfices tirés de cette analyse).

La prospérité attend les sociétés qui vont pouvoir se situer au cœur des flux d'informations et collecter des données. Les exploiter efficacement exige des compétences techniques et une bonne dose d'imagination – un état d'esprit orienté big data. Mais l'essentiel de la valeur revient aux détenteurs de données. Et il arrive que ce ne soient pas les informations clairement visibles qui soient les plus importantes, mais les traces numériques nées des interactions des internautes avec les informations. Une société ingénieuse pourra alors les utiliser pour améliorer un service existant ou en lancer un complètement nouveau.

En même temps, les big data nous exposent à des risques énormes. Elles rendent inefficaces les principaux mécanismes techniques et les instruments juridiques de protection de la vie privée en service actuellement. Par le passé, on connaissait bien ce qui constituait l'information et permettait d'identifier un individu – son nom, numéro de Sécurité sociale, le registre fiscal, etc. – des éléments relativement faciles à protéger. Aujourd'hui, même les données les plus anodines peuvent révéler l'identité d'un individu si un collecteur de données en a amassé un nombre suffisant. Anonymiser les données ou les cacher dans la masse ne fonctionne plus. S'il est ciblé, un individu peut voir l'intrusion dans sa vie privée devenir bien plus importante que jadis : les autorités veulent en effet non seulement obtenir le maximum d'informations sur lui, mais aussi connaître le plus grand nombre de ses relations, de ses contacts et de ses interactions.

Outre la protection de la vie privée, les big data créent un nouveau motif d'inquiétude : le risque que nous jugions les gens non pas simplement sur la base de leur comportement réel mais sur des intentions que révèlent les données. Au fur et à mesure que les prévisions issues de big data gagneront en exactitude, la société pourra s'en servir pour sanctionner des personnes pour un acte qu'elles prédisent mais qui n'a pas encore eu lieu. Des prédictions impossibles à réfuter, et dont les personnes accusées ne

peuvent jamais se disculper. Une sanction décidée sur cette base va à l'encontre du concept du libre arbitre et nie la possibilité, même ténue, qu'une personne choisisse une voie différente. Lorsque la société détermine la responsabilité individuelle (et prescrit la sanction), elle se doit de considérer la volonté humaine comme inviolable. Nous devons conserver la possibilité de façonner l'avenir selon notre libre conception. Autrement, les big data auront corrompu ce qui fait l'essence même de l'humanité : la pensée rationnelle et le libre arbitre.

Il n'existe pas de méthode infaillible pour se préparer complètement au monde des big data ; il nous faudra établir de nouveaux principes pour nous régir. Une série de changements importants dans nos pratiques peut aider la société durant le processus de familiarisation avec la nature des big data et leurs insuffisances. Nous devons protéger la vie privée à travers une exploitation responsable des données par leurs utilisateurs. Dans un monde de prédictions, il est vital que la volonté humaine conserve son caractère sacré et que nous sauvegardions non seulement la possibilité de faire un choix moral mais aussi d'assumer la responsabilité de nos actes. La société doit concevoir des mesures de protection pour permettre à une nouvelle classe d'experts, les « algorithmistes », d'évaluer les analyses de big data – de manière que ce monde devenu moins aléatoire grâce aux big data ne se transforme pas en une boîte noire, où l'on aurait simplement remplacé une forme de l'inconnaissable par une autre.

Les ensembles de données deviendront essentiels dans la compréhension et la résolution de bon nombre des problèmes mondiaux urgents. Pour la lutte contre le changement climatique, il faut analyser les données sur la pollution, comprendre comment concentrer nos efforts de la meilleure manière et trouver les moyens de limiter les problèmes. Les capteurs implantés dans le monde entier, notamment ceux intégrés dans les smartphones, fournissent une mine de données qui vont nous permettre de modéliser le réchauffement planétaire de manière plus détaillée. Entre-temps, l'automatisation de certaines tâches va permettre d'améliorer les soins de santé et d'en faire baisser les coûts, en particulier pour les populations pauvres dans le monde : par exemple les biopsies ou bien la détection d'infections avant l'apparition complète des symptômes, dont on a aujourd'hui l'impression qu'elles nécessitent l'intervention humaine, pourront être effectuées par ordinateur.

Les big data servent déjà au développement économique et à la prévention des conflits. En analysant comment se déplacent les utilisateurs de téléphones portables, on a pu voir que, dans certains bidonvilles africains, s'implantaient des communautés entrepreneuriales qui développaient des activités économiques, d'autres où le contexte était mûr pour des affrontements ethniques et elles ont même permis de prévoir les crises liées au problème des réfugiés. Les utilisations des big data se multiplieront au fur et à mesure que cette technologie sera appliquée à des aspects encore plus nombreux de notre vie.

Si les big data nous aident à améliorer ce que nous faisons déjà, elles nous permettent aussi de réaliser des choses complètement nouvelles. Elles n'en sont pas pour autant une baguette magique ! Elles ne rétabliront pas la paix dans le monde, elles n'éradiqueront pas la pauvreté ni ne feront apparaître un nouveau Picasso ! Les big data sont certes incapables de faire un enfant – mais elles peuvent en revanche sauver la vie de prématurés. À terme, nous devons nous attendre à les voir utilisées dans quasiment tous les aspects de la vie, au point que leur absence susciterait en nous une légère inquiétude, comme lorsque nous attendons d'un médecin qu'il nous prescrive la radiographie plus à même qu'un examen clinique de révéler nos problèmes.

L'utilisation des big data se généralisera, et elle pourrait bien influencer alors sur notre façon d'envisager l'avenir. Il y a environ cinq cents ans, l'humanité a connu un changement profond dans la perception du temps, un des éléments importants dans l'évolution en Europe vers une société éclairée, plutôt séculière et s'appuyant sur les sciences. Auparavant, le temps était appréhendé, tout comme la vie, dans sa dimension cyclique. Toutes les années et tous les jours ressemblaient à ceux qui les avaient précédés, et la fin de vie elle-même était semblable à son commencement, avec des vieillards retombant en enfance. Le temps fut ensuite perçu comme linéaire – comme une succession de jours où l'on pouvait façonner le monde et influencer sur la trajectoire de la vie. Alors qu'en des époques plus anciennes, passé, présent et futur étaient soudés, l'humanité s'est désormais découvert un passé vers lequel très distinctement se retourner, un futur à attendre avec intérêt et un présent à façonner.

Au moment où le présent est devenu modelable, le futur, jusque-là prévisible, s'est offert comme une vaste page blanche où chacun pouvait désormais faire figurer ses efforts et ses valeurs. L'un des traits marquants

des temps modernes est que l'homme a le sentiment d'être maître de son destin ; cette attitude nous distingue de nos ancêtres, pour qui une certaine forme de déterminisme était la norme. Et voici que les prédictions de big data nous font apparaître un futur qui n'est plus une page blanche à remplir librement, mais un tableau où sont déjà esquissées, peu visibles par nous, les lignes de notre avenir parfaitement discernables par ceux qui disposent de la technologie. Cela semble bien réduire notre capacité à façonner notre destinée ! La potentialité est sacrifiée sur l'autel de la probabilité.

En même temps, les big data pourraient signifier que nous sommes à jamais prisonniers de nos actions antérieures, utilisables à notre encontre par des systèmes qui prétendent prédire notre comportement futur : nous ne pouvons jamais échapper à ce qui s'est produit antérieurement. « Le passé est prologue », a écrit Shakespeare. C'est cette idée que les big data consacrent sous forme d'algorithmes, et cela, dans un sens positif ou négatif. Un monde de prédictions éteindra-t-il l'enthousiasme qui anime notre désir de marquer le monde de notre empreinte humaine et qui nous fait saluer le lever du soleil ?

C'est en fait le contraire qui est le plus probable ; en effet, connaître à l'avance des actions à venir permet de prendre les mesures correctives de façon à prévenir les problèmes ou améliorer les résultats. Il serait possible de repérer, avant les examens de fin d'année, les étudiants dont le niveau commence à faiblir. De détecter des cancers à leur tout début et de les traiter avant que la maladie n'apparaisse. Nous verrions avec quelle probabilité une adolescente risque d'avoir une grossesse non désirée ou que quelqu'un s'engage sur la voie du crime. Nous pourrions alors intervenir pour modifier, dans la mesure du possible, l'issue annoncée par la prédiction. Nous empêcherions que ne surviennent des incendies meurtriers dans les habitations surpeuplées à New York en sachant quels bâtiments inspecter en priorité.

Rien n'est prédestiné, puisqu'il nous est toujours possible de réagir face à l'information que nous recevons. Les prédictions des big data ne sont pas gravées dans le marbre – elles ne sont que des issues probables, et cela signifie que si nous voulons les modifier, nous le pouvons. Nous saurons déterminer la meilleure attitude à adopter pour envisager l'avenir et en être le maître, tout comme Maury a mis au jour des voies maritimes naturelles et ainsi amélioré la navigation. Et pour accomplir de telles choses, nous

n'aurons pas besoin de comprendre la nature du cosmos ni de prouver l'existence des dieux – les big data suffiront.

Des volumes de données encore plus grands

Avec le changement que les big data vont apporter dans notre vie – optimiser, améliorer, augmenter l'efficacité et tirer des bénéfices de tout cela –, quelle place va-t-il rester à l'intuition, à la foi, au doute et à l'originalité ?

S'il y a un enseignement à tirer des big data, c'est que se contenter d'agir au mieux et d'apporter des améliorations – sans chercher à approfondir – est souvent tout à fait suffisant. Il est sage de continuer à adopter ce comportement. Même si vous ne savez pas pourquoi vos efforts sont fructueux, ils donnent de meilleurs résultats que si vous vous abstenez de les poursuivre. Flowers et ses « gosses » à New York n'incarnent sans doute pas l'esprit éclairé que les sages présentent, mais pourtant ils sauvent des vies.

Les big data ne sont pas un monde glacé d'algorithmes et d'automates. Les êtres humains que nous sommes ont un rôle essentiel à jouer, avec toutes leurs faiblesses, leurs idées fausses et leurs erreurs, car ce sont là des traits qui vont de pair avec la créativité, l'instinct et le génie propres aux humains. Ce sont les mêmes processus mentaux désordonnés qui nous amènent parfois à vivre une situation humiliante ou à nous faire faire fausse route, ce qui peut avoir des retombées positives et révéler à nous-mêmes ce qu'il y a en nous d'excellent. Cela suggère que, tout comme nous apprenons à faire avec des données déstructurées parce qu'elles servent un objectif plus grand, l'inexactitude qui fait partie intégrante de la nature humaine doit être tolérée. Après tout, le désordre est une caractéristique essentielle du monde et de notre esprit, et dans les deux cas, ce n'est qu'en l'admettant et en l'exploitant que nous en tirons avantage.

Dans un monde où les données éclairent les décisions, de quelle latitude disposera-t-on pour agir selon son intuition et pour aller dans un sens différent de celui qu'indiquent les faits ? Si tout le monde fait appel aux données et utilise les outils des big data, ce qui créera la différence essentielle viendra sans doute de l'imprévisibilité : cette part de la dimension humaine qui relève de l'instinct, de la prise de risque, de l'accident et de l'erreur.

Il faudrait veiller tout particulièrement à ce que l'être humain conserve sa place et sache sauvegarder ce qui doit revenir à l'intuition, au bon sens et à la chance. Et ce, pour empêcher qu'il ne soit supplanté par les données et par les réponses automatiques des machines. La dimension la plus grande chez les êtres humains est précisément ce que les algorithmes et les puces de silicium ne révèlent pas et ne peuvent révéler parce que cela n'est pas saisissable dans les données ! Il ne s'agit pas de « ce qui est », mais de « ce qui n'est pas » : l'espace vide, les fissures du trottoir, l'informulé, le non encore conçu.

Cela a d'importantes implications pour la notion de progrès. Les big data nous permettent d'expérimenter plus rapidement et d'explorer de plus nombreuses pistes. Ces avantages devraient générer plus d'innovation. Mais ce que les données ne disent pas, c'est ce que produit l'étincelle de l'invention : il y a en effet une chose qu'aucun volume de données ne pourra jamais corroborer, c'est celle qui n'a pas encore été inventée. Si Henry Ford, par exemple, avait interrogé les algorithmes des big data pour connaître le souhait de ses clients, il aurait eu pour réponse, selon sa célèbre boutade : « un cheval plus rapide ». Dans un monde de big data, il faudrait que nos qualités humaines de créativité et d'intuition, et nos ambitions intellectuelles soient d'autant plus favorisées que c'est de notre ingéniosité que jaillit le progrès.

Les big data sont à la fois une ressource et un outil, destinées à informer plutôt qu'à expliquer : elles peuvent nous faire découvrir des choses, mais néanmoins provoquer des malentendus, selon qu'elles sont utilisées à bon ou à mauvais escient. Quand bien même nous serions éblouis par le pouvoir des big data, nous ne devons jamais nous laisser aveugler par leur éclatante séduction ni oublier leurs imperfections.

Nous ne pourrions jamais recueillir, stocker ni traiter la totalité des informations dans le monde – l'ultime N = tous – dans l'état actuel de nos technologies. Par exemple, le laboratoire de physique des particules du CERN en Suisse recueille moins de 0,1 % des informations lors des expériences qui y sont menées – le reste, vraisemblablement sans aucune utilité, disparaît dans la nature. Mais ceci n'est pas nouveau. Depuis toujours, la société s'est heurtée aux limites des outils de mesure de la réalité, du compas et du sextant jusqu'au GPS aujourd'hui en passant par le télescope et le radar. Demain, nos outils seront deux fois, dix fois voire mille fois plus puissants qu'ils ne le sont aujourd'hui, et nos connaissances

actuelles nous sembleront alors minuscules. Le monde présent des big data nous paraîtra sous peu aussi dépassé que les quatre kilo-octets de mémoire vive de l'ordinateur de bord d'Apollo 11.

Mais ce que nous sommes capables de collecter et de traiter ne sera jamais qu'une minuscule fraction des informations existant dans le monde. Un simulacre de la réalité tout comme les ombres projetées sur les murs de la caverne de Platon. Il ne sera jamais possible d'obtenir des informations parfaites et c'est pourquoi nos prédictions sont intrinsèquement faillibles. Non pas qu'elles soient fausses, mais simplement parce qu'elles sont toujours incomplètes. Cela ne remet nullement en question les découvertes que nous offrent les big data, mais ces dernières doivent être prises pour ce qu'elles sont : un outil qui nous apporte des réponses non définitives, mais suffisantes en attendant que se présentent de meilleures méthodes et donc de meilleures réponses. Cela nous exhorte à utiliser cet outil avec beaucoup d'humilité... et d'humanité.

[a.](#) Prêts hypothécaires à risque, à l'origine de la crise financière mondiale.

Notes

Chapitre 1. Aujourd'hui

1. Google Suivi de la grippe – Jeremy Ginsburg *et al.*, « Detecting Influenza Epidemics Using Search Engine Query Data », *Nature* 457 (2009), pp. 1012-1014 (<http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html>).

2. Étude complémentaire sur Google Suivi de la grippe – A. F. Dugas *et al.*, « Google FluTrends : Correlation with Emergency Department Influenza Rates and Crowding Metrics », CID Advanced Access (8 janvier 2012) ; DOI 10.1093/cid/cir883.

3. Achat de billets d'avion, Farecast – Informations extraites de Kenneth Cukier, « Data, Data Everywhere », *The Economist* dossier spécial, 27 février 2010, pp. 1-14, et d'entretiens avec Etzioni entre 2010 et 2012. Projet Hamlet d'Etzioni – Oren Etzioni, C. A. Knoblock, R. Tuchinda et A. Yates, « To Buy or Not to Buy : Mining Airfare Data to Minimize Ticket Purchase Price », SIGKDD '03, 24-27 août 2003 (<http://knight.cis.temple.edu/~yates/papers/hamlet-kdd03.pdf>).

4. Prix payé par Microsoft pour l'acquisition de Farecast – Information provenant de communiqués de presse, notamment « Secret Farecast Buyer Is Microsoft », *Seattlepi.com*, 17 avril 2008 (<http://blog.seattlepi.com/venture/2008/04/17/secret-farecast-buyer-is-microsoft/?source=myspi>).

5. Une définition possible des big data – Le débat sur l'origine du terme « big data » et sur la meilleure définition du terme est très vif mais il n'a pas abouti. Les deux mots accolés sont apparus occasionnellement ces dernières décennies. Doug Laney, de la société de recherche Gartner, a énoncé en 2001 dans un rapport de recherche les « trois V » de big data (volume,

vitesse et diversité – *variety*, en anglais), définition utile à l'époque mais néanmoins imparfaite.

6. Astronomie et séquençage de l'ADN – Cukier, « Data, Data Everywhere ». Des milliards d'actions négociées – Rita Nazareth and Julia Leite, « Stock Trading in U.S. Falls to Lowest Level Since 2008 », *Bloomberg*, 13 août 2012 (<http://www.bloomberg.com/news/2012-08-13/stock-trading-in-u-s-hits-lowest-level-since-2008-as-vix-falls.html>).

7. 24 pétaoctets traités quotidiennement par Google – Thomas H. Davenport, Paul Barth et Randy Bean, « How “Big Data” Is Different », *Sloan Review*, 30 juillet 2012, pp. 43-46 (<http://sloanreview.mit.edu/the-magazine/2012-fall/54104/how-big-data-is-different/>). Statistiques de Facebook – prospectus d'introduction en Bourse de Facebook, « Form S-1 Registration Statement », U.S. Securities and Exchange Commission, 1^{er} février 2012 (<http://sec.gov/Archives/edgar/data/1326801/000119312512034517/d287954ds1.htm>). YouTube stats – Larry Page, « Update from the CEO », Google, avril 2012 (<http://investor.google.com/corporate/2012/ceo-letter.html>). Nombre de tweets – Tomio Geron, « Twitter's Dick Costolo : Twitter MobileAd Revenue Beats Desktop on Some Days », *Forbes*, 6 juin 2012 (<http://www.forbes.com/sites/tomiogeron/2012/06/06/twitters-dick-costolo-mobile-ad-revenue-beats-desktop-on-some-days/>). Informations sur le volume de données – Martin Hilbert and Priscilla López, « The World's Technological Capacity to Store, Communicate, and Compute Information », *Science*, 1 (avril 2011), pp. 60-65 ; Martin Hilbert et Priscilla López, « How to Measure the World's Technological Capacity to Communicate, Store and Compute Information ? », *International Journal of Communication* 6 (2012), pp. 1042-55 (<http://www.ijoc.org/ojs/index.php/ijoc/article/viewFile/1562/742>).

8. Estimation de la quantité d'informations stockées à 2013 – Entretien de Cukier avec Hilbert, 2012.

9. Presse à imprimer et huit millions de livres ; nombre supérieur à celui des livres produits depuis la fondation de Constantinople – Elizabeth L. Eisenstein, *The Printing Revolution in Early Modern Europe*, Canto/Cambridge University Press, 1993, pp. 13-14. Analogie présentée par Peter Norvig – Extrait de conférences données par Norvig et fondées sur l'article : A. Halevy, P. Norvig et F. Pereira, « The Unreasonable Effectiveness of Data^a », *IEEE Intelligent Systems*, mars/avril 2009, pp. 8-12

(http://www.computer.org/portal/cms_docs_intelligent/intelligent/homepage/2009/x2exp.pdf). (Remarque : le titre fait écho à l'article d'Eugene Wigner, « The Unreasonable Effectiveness of Mathematics in the Natural Sciences », dans lequel il examine la raison pour laquelle les lois physiques peuvent être tout à fait bien énoncées à l'aide de formules mathématiques, mais les sciences sociales se prêtent mal à l'utilisation de ces formules bien nettes. Voir E. Wigner, « The Unreasonable Effectiveness of Mathematics in the Natural Sciences^b », *Communications on Pure and Applied Mathematics* 13, n° 1 (1960), pp. 1-14.) Une des conférences de Norvig au sujet de l'article est celle intitulée « Peter Norvig – L'efficacité déraisonnable des données », effectuée à l'Université de la Colombie britannique, YouTube, 23 septembre 2010 (<http://www.youtube.com/watch?v=yvDCzhbjYWs>). Au sujet de l'influence de la taille physique sur la loi physique opérante (bien que cela ne soit pas entièrement correct), la référence souvent citée est J. B. S. Haldane, « On Being the Right Size », *Harper's Magazine*, mars 1926 (<http://harpers.org/archive/1926/03/on-being-the-right-size>). Propos de Picasso au sujet des images de Lascaux – David Whitehouse, « UK Science Shows Cave Art Developed Early », *BBC News Online*, 3 octobre 2001 (<http://news.bbc.co.uk/1/hi/sci/tech/1577421.stm>).

Chapitre 2. Plus

10. Citation de Jeff Jonas – Conversation avec Jonas, décembre 2010, Paris.

11. Histoire du recensement américain – Bureau américain du recensement, récit sur la « Machine de Hollerith » en ligne. (http://www.census.gov/history/www/innovations/technology/the_hollerith_tabulator.html).

12. Contribution de Neyman – William Kruskal et Frederick Mosteller, « Representative Sampling, IV : The History of the Concept in Statistics, 1895-1939 », *International Statistical Review* 48 (1980), pp. 169-195 et pp. 187-188. Le célèbre article de Jerzy Neyman est intitulé « On the Two Different Aspects of the Representative Method : The Method of Stratified Sampling and the Method of Purposive Selection », *Journal of the Royal Statistical Society* 97, n° 4 (1934), pp. 558-625. Un échantillon de 1 100 observations est suffisant – Earl Babbie, *Practice of Social Research* (12^e éd. 2010), pp. 204-207.

13. L'effet des téléphones cellulaires – « Estimating the Cellphone Effect », 20 septembre 2008

(<http://www.fivethirtyeight.com/2008/09/estimating-cellphone-effect-22-points.html>) ; pour en savoir plus sur les biais dans les sondages et avoir d'autres informations statistiques, voir Nate Silver, *The Signal and the Noise : Why So Many Predictions Fail – But Some Don't* (Penguin, 2012).

14. Séquençage de l'ADN de Steve Jobs – Walter Isaacson, *Steve Jobs* (Simon and Schuster, 2011), pp. 550-551.

15. Prédictions de Google Suivi de la grippe au niveau des villes – Dugas *et al.*, « Google FluTrends ». Etzioni sur les données temporelles – Entretien avec Cukier, octobre 2011.

16. Citation de John Kunze – Jonathan Rosenthal, « Special Report : International Banking », *The Economist*, 19 mai 2012, pp. 7-8. Trucage de matchs de sumo – Mark Duggan et Steven D. Levitt, « Winning Isn't Everything : Corruption in Sumo Wrestling », *American Economic Review* 92 (2002), pp. 1594-1605 (<http://pricetheory.uchicago.edu/levitt/Papers/DugganLevitt2002.pdf>).

17. 11 millions de rayons lumineux du Lytro – site Web de Lytro (<http://www.lytro.com>).

18. Remplacement de l'échantillonnage dans les sciences sociales – Mike Savage et Roger Burrows, « The Coming Crisis of Empirical Sociology », *Sociology* 41 (2007), pp. 885-899. Sur l'analyse de données à grande échelle provenant d'un opérateur de téléphonie mobile – J. P. Onnela *et al.*, « Structure and Tie Strengths in Mobile Communication Networks », *Proceedings of the National Academy of Sciences of the United States of America* (PNAS) 104 (mai 2007), pp. 7332-7336 (<http://nd.edu/~dddas/Papers/PNAS0610245104v1.pdf>).

Chapitre 3. Désordre

19. Crosby – Alfred W. Crosby, *The Measure of Reality : Quantification and Western Society, 1250-1600* (Cambridge University Press, 1997). Citations de Kelvin et Bacon – Ces aphorismes sont généralement attribués aux deux hommes, même si l'expression figurant réellement dans leurs œuvres respectives est légèrement différente. Pour Kelvin, elle fait partie d'une citation plus longue sur la mesure, extraite de sa conférence « Electrical Units of Measurement » (1883). Pour Bacon, elle est considérée

comme une traduction libre d'une citation en latin dans *Meditationes Sacrae* (1597).

20. Les nombreuses appellations d'IBM – DJ Patil, « Data Jujitsu : The Art of Turning Data into Product », *O'Reilly Media*, juillet 2012 (<http://oreillynnet.com/oreilly/data/radarreports/data-jujitsu.csp?cmp=tw-strata-books-data-products>).

21. 30 000 transactions par seconde à la Bourse de New York – Colin Clark, « Improving Speed and Transparency of Market Data », billet sur le blog NYSE EURONEXT, 9 janvier 2011 (<http://exchanges.nyx.com/cclark/improving-speed-and-transparency-market-data>). Idée que « $2 + 2 = 3,9$ » – Brian Hopkins et Boris Evelson, « Expand Your Digital Horizon with Big Data », Forrester, 30 septembre 2011. Meilleures performances des algorithmes – President's Council of Advisors on Science and Technology, « Report to the President and Congress, Designing a Digital Future : Federally Funded Research and Development in Networking and Information Technology », décembre 2010, p. 71 (<http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf>).

22. Tables de fins de partie aux échecs – Table de finale la plus complète accessible au public, l'ensemble de tables de finales de Nalimov (du nom de leur créateur), couvre toutes les finales de six pièces ou moins. Sa taille dépasse les sept téraoctets, et la compression des données y est extrêmement difficile. Voir E. V. Nalimov, G. McC. Haworth et E. A. Heinz, « Space-efficient Indexing of Chess Endgame Tables », *ICGA Journal* 23, n° 3 (2000), pp. 148-162. Microsoft et performance des algorithmes – Michele Banko et Eric Brill, « Scaling to Very Very Large Corpora for Natural Language Disambiguation », *Microsoft Research*, 2001, p. 3 (<http://acl.ldc.upenn.edu/P/P01/P01-1005.pdf>).

23. Démonstration, mots et citation d'IBM – IBM, « 701 Translator », communiqué de presse, archives IBM, 8 janvier 1954 (http://www-03.ibm.com/ibm/history/exhibits/701/701_translator.html). Voir aussi John Hutchins, « The First Public Demonstration of Machine Translation : The Georgetown-IBM System, 7 janvier 1954 », novembre 2005 (<http://www.hutchinsweb.me.uk/GU-IBM-2005.pdf>). Le projet Candide d'IBM – Adam L. Berger *et al.*, « The Candide System for Machine Translation », *Proceedings of the 1994 ARPA Workshop on Human*

Language Technology, 1994 (<http://aclweb.org/anthology-new/H/H94/H94-1100.pdf>). Histoire de la traduction automatique – Yorick Wilks, *Machine Translation : Its Scope and Limits* (Springer, 2008), p. 107.

24. Les millions de textes du projet Candide par rapport aux milliards de textes de Google – Entretien d'Och avec Cukier, décembre 2009. Le corpus de 95 milliards de phrases de Google – Alex Franz et Thorsten Brants, « All Our N-gram are Belong to You », billet sur le blog de Google, 3 août 2006 (<http://googleresearch.blogspot.co.uk/2006/08/all-our-n-gram-are-belong-to-you.html>).

25. Brown corpus et le billion de mots de Google – Halevy, Norvig et Pereira, « The Unreasonable Effectiveness of Data ». Citation d'un article dont Norvig est coauteur – *Ibid.*

26. Corrosion de la tuyauterie d'une raffinerie de BP et environnement hostile pour les capteurs sans fil – Jaclyn Clarabut, « Operations Making Sense of Corrosion », *BP Magazine*, 2 (2011) (http://www.bp.com/liveassets/bp_internet/globalbp/globalbp_uk_english/reports_and_publications/bp_magazine/STAGING/local_assets/pdf/BP_Magazine_2011_issue2_text.pdf). Difficulté d'obtenir des relevés de données exacts, évoquée dans Cukier, « Data, Data, Everywhere ». Le système n'est manifestement pas infallible : un incendie à la raffinerie Cherry Point de BP en février 2012 a été imputé à la corrosion d'une canalisation.

27. Billion Prices Project – Entretien des cofondateurs avec Cukier, octobre 2012. Également, James Surowiecki, « A Billion Prices Now », *The New Yorker*, 30 mai 2011 ; les données et les détails sont disponibles sur le site Web du projet (<http://bpp.mit.edu/>) ; Annie Lowrey, « Economists' Programs Are Beating U.S. at Tracking Inflation », *Washington Post*, 25 décembre 2010 (<http://www.washingtonpost.com/wp-dyn/content/article/2010/12/25/AR2010122502600.html>).

28. Supériorité des chiffres fournis par Price Stats par rapport aux statistiques nationales – « Official Statistics : Don't Lie to Me, Argentina », *The Economist*, 25 février 2012 (<http://www.economist.com/node/21548242>). Nombre de photos sur Flickr – sur le site Web de Flickr (<http://www.flickr.com>). Difficulté de catégoriser les informations – Voir David Weinberger, *Everything Is Miscellaneous : The Power of the New Digital Disorder* (Times Books, 2007).

29. Pat Helland – Pat Helland, « If You Have Too Much Data Then "Good Enough" Is Good Enough », *Communications of the ACM*,

juin 2011, pp. 40-41. Dans le milieu des bases de données, le débat est vif, concernant les modèles et les concepts les plus adaptés pour répondre aux exigences en termes de big data. Helland représente le camp qui prône la rupture radicale avec les outils utilisés dans le passé. Michael Rys de Microsoft, in « Scalable SQL », *Communications of the ACM*, juin 2011, p. 48, avance quant à lui que des versions bien adaptées des outils existants peuvent fonctionner correctement.

30. Utilisation de Hadoop par Visa – Cukier, « Data, data everywhere ».

31. Seulement 5 % des informations sont des données structurées – Abhishek Mehta, « Big Data : Powering the Next Industrial Revolution », Tableau Software White Paper, 2011 (<http://www.tableausoftware.com/learn/whitepapers/big-data-revolution>).

Chapitre 4. Corrélation

32. L'histoire de Linden et de la « voix d'Amazon » – Entretien de Linden avec Cukier, mars 2012. Article du *Wall Street Journal* sur les critiques d'Amazon – Cité dans James Marcus, *Amazonia : Five Years at the Epicenter of the Dot.Com Juggernaut* (New Press, 2004), p. 128.

33. Citation de Marcus – Marcus, *Amazonia*, p. 199.

34. Les recommandations, source d'un tiers du chiffre d'affaires d'Amazon – Ce chiffre n'a jamais été officiellement confirmé par la société mais il a été publié dans de nombreux rapports d'analystes et articles dans les médias, notamment « Building with Big Data : The Data Revolution Is Changing the Landscape of Business », *The Economist*, 26 mai 2011 (<http://www.economist.com/node/18741392>). Ce chiffre a également été évoqué par deux anciens directeurs d'Amazon lors d'entretiens avec Cukier. Informations sur les prix de Netflix – Xavier Amatriain et Justin Basilico, « Netflix Recommendations : Beyond the 5 stars (Part 1) », blog de Netflix, 6 avril 2012.

35. « Trompés par le hasard » – Nassim Nicholas Taleb, *Le Hasard sauvage. Comment la chance nous trompe* (Les Belles Lettres, Paris, 2009) ; pour en savoir plus, voir Nassim Nicholas Taleb, *Le Cygne noir. La puissance de l'imprévisible* (Les Belles Lettres, Paris, 2008).

36. Walmart et Pop-Tarts – Constance L. Hays, « What WalMart Knows About Customers' Habits », *New York Times*, 14 novembre 2004 (<http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html>).

37. Exemples de modèles prédictifs de FICO, Experian et Equifax – Scott Thurm, « Next Frontier in Credit Scores : Predicting Personal Behavior », *Wall Street Journal*, 27 octobre 2011 (<http://online.wsj.com/article/SB10001424052970203687504576655182086300912.html>).

38. Modèles prédictifs d'Aviva – Leslie Scism et Mark Maremont, « Insurers Test Data Profiles to Identify Risky Clients », *Wall Street Journal*, 19 novembre 2010 (<http://online.wsj.com/article/SB10001424052748704648604575620750998072986.html>). Voir aussi Leslie Scism et Mark Maremont, « Inside Deloitte's Life-Insurance Assessment Technology », *Wall Street Journal*, 19 novembre 2010 (<http://online.wsj.com/article/SB10001424052748704104104575622531084755588.html>). Voir aussi Howard Mills, « Analytics : Turning Data into Dollars », *Forward Focus*, décembre 2011 (http://www.deloitte.com/assets/Dcom-UnitedStates/Local%20Assets/Documents/FSI/US_FSI_Forward%20Focus_Analytics_Turning%20data%20into%20dollars_120711.pdf). Exemple de Target et de l'adolescente enceinte – Charles Duhigg, « How Companies Learn Your Secrets », *New York Times*, 16 février 2012 (<http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html>).

L'article est adapté du livre de Duhigg, *Le Pouvoir des habitudes. Changer un rien pour tout changer* (Saint-Simon, 2013) ; Target a déclaré qu'il y avait des inexactitudes dans la présentation par les médias de ses activités, mais refuse de dire lesquelles. Interrogé sur le sujet pour les besoins du livre, un porte-parole de Target a répondu : « Le but est d'utiliser les données pour améliorer la relation des clients avec Target. Nos clients attendent de nous une valeur accrue, des offres mieux ciblées et souhaitent obtenir un service client hautement satisfaisant. Comme de nombreuses sociétés, nous nous servons d'outils qui nous aident à comprendre les habitudes d'achat et les goûts de nos clients de manière à ce que nos offres et nos promotions leur correspondent le mieux possible. Nous prenons très au sérieux la responsabilité que nous avons de préserver la confiance placée en nous par nos clients. Nous avons pour cela opté pour deux initiatives : une politique complète de protection de la vie privée présentée au public sur notre site Web Target.com et une formation régulière des membres de notre équipe sur les méthodes de sécurisation des informations de nos clients. »

[39.](#) Utilisation de l'analyse par UPS – Entretiens de Jack Levis avec Cukier, 2012.

[40.](#) Prématurés – D'après des entretiens avec McGregor en 2010 et 2012. Voir aussi Carolyn McGregor, Christina Catley, Andrew James et James Padbury, « Next Generation Neonatal Health Informatics with Artemis », dans European Federation for Medical Informatics, *User Centred Networked HealthCare*, ed. A. Moen *et al.* (IOS Press, 2011), p. 117. Une partie de la documentation provient de Cukier, « Data, Data, Everywhere ».

[41.](#) Corrélation entre le bonheur et le revenu – R. Inglehart et H.-D. Klingemann, *Genes, Culture and Happiness* (MIT Press, 2000).

[42.](#) Rougeole et dépenses de santé, et nouveaux outils d'analyse de corrélations non linéaires – David Reshef *et al.*, « Detecting Novel Associations in Large Data Sets », *Science* 334 (2011), pp. 1518-24.

[43.](#) Kahneman – Daniel Kahneman, *Système 1, système 2. Les deux vitesses de la pensée* (Flammarion, 2012).

[44.](#) Pasteur – Pour les lecteurs intéressés par l'influence plus large exercée par Pasteur sur notre perception des choses, nous suggérons l'ouvrage de Bruno Latour, *Pasteur. Guerre et paix des microbes*, La Découverte, Paris, 2011. Risque d'être contaminé par la rage – Melanie Di Quinzio et Anne McCarthy, « Rabies Risk Among Travellers », *CMAJ* 178, n° 5 (2008), p. 567.

[45.](#) La causalité peut rarement être prouvée – L'informaticien Judea Pearl, lauréat du prix Turing, a élaboré une méthode de représentation de la dynamique causale ; sans constituer une preuve formelle, cette méthode propose une approche pragmatique pour analyser les relations causales ; voir Judea Pearl, *Causality : Models, Reasoning and Inference* (Cambridge University Press, 2009).

[46.](#) Exemple de la voiture orange – Quentin Hardy, « Bizarre Insights from Big Data », *nytimes.com*, 28 mars 2012 (<http://bits.blogs.nytimes.com/2012/03/28/bizarre-insights-from-big-data/>) ; et Kaggle, « Momchil Georgiev Shares His Chromatic Insight from Don't Get Kicked », billet sur son blog, 2 février 2012 (<http://blog.kaggle.com/2012/02/02/momchil-georgiev-shares-his-chromatic-insight-from-dont-get-kicked/>).

[47.](#) Poids des couvercles de trous d'homme, nombre d'explosions et hauteur de projection des morceaux – Rachel Ehrenberg, « Predicting the

Next Deadly Manhole Explosion », *Wired*, 7 juillet 2010 (<http://www.wired.com/wiredscience/2010/07/manhole-explosions>).

Collaboration de Con Edison avec les statisticiens de l'Université de Columbia – Le public profane trouvera une description de ce cas dans Cynthia Rudin *et al.*, « 21st-Century Data Miners Meet 19th-Century Electrical Cables », *Computer*, juin 2011, pp. 103-105. Des descriptions techniques de ce travail figurent dans les articles scientifiques de Rudin et de ses collaborateurs sur leurs sites Web respectifs, en particulier Cynthia Rudin *et al.*, « Machine Learning for the New York City Power Grid », *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, n° 2 (2012), pp. 328-345 (<http://hdl.handle.net/1721.1/68634>).

48. Données désordonnées relatives au terme « coffre de branchement » – Cette liste est extraite de Rudin *et al.*, « 21st-Century Data Miners Meet 19th-Century Electrical Cables ». Citation de Rudin – Entretien avec Cukier, mars 2012.

49. Opinion d'Anderson – Chris Anderson, « The End of Theory : The Data Deluge Makes the Scientific Method Obsolete », *Wired*, juin 2008 (http://www.wired.com/science/discoveries/magazine/16-07/pb_theory/).

50. Retour d'Anderson sur ses affirmations – National Public Radio, « Search and Destroy », 18 juillet 2008 (<http://www.onthemedial.org/2008/jul/18/search-and-destroy/transcript>).

51. Influence des choix sur notre analyse – Danah Boyd et Kate Crawford. « Six Provocations for Big Data », article présenté lors du Symposium organisé à l'Oxford Internet Institute « A Decade in Internet Time : Symposium on the Dynamics of the Internet and Society », 21 septembre 2011 (<http://ssrn.com/abstract=1926431>).

Chapitre 5. Mise en données

52. Les détails sur la vie de Maury sont extraits de nombreux ouvrages écrits par lui et sur lui, notamment : Chester G. Hearn, *Tracks in the Sea : Matthew Fontaine Maury and the Mapping of the Oceans* (International Marine/McGraw-Hill, 2002) ; Janice Beaty, *Seeker of Seaways : A Life of Matthew Fontaine Maury, Pioneer Oceanographer* (Pantheon Books, 1966) ; Charles Lee Lewis, *Matthew Fontaine Maury : The Pathfinder of the Seas* (U.S. Naval Institute, 1927).

(<http://archive.org/details/matthewfontainem00lewi>) ; et Matthew Fontaine Maury, *The Physical Geography of the Sea* (Harper, 1855).

53. Citations de Maury – Extraites de Maury, *Physical Geography of the Sea*, « Introduction », pp. xii, vi.

54. Données sur les sièges de voiture – Nikkei, « Car Seat of Near Future IDs Driver's Backside », 14 décembre 2011.

55. Quantifier le monde – Une grande partie de la réflexion des auteurs sur l'histoire de la mise en données a été inspirée par Crosby, *The Measure of Reality*.

56. Absence de contact des Européens avec les bouliers – *Ibid.*, 112. Calculer plus rapidement avec les chiffres arabes – Alexander Murray, *Reason and Society in the Middle Ages* (Oxford University Press, 1978), p. 166.

57. Nombre total de livres publiés et étude de chercheurs d'Harvard sur le projet de scannage de livres de Google – Jean-Baptiste Michel *et al.*, « Quantitative Analysis of Culture Using Millions of Digitized Books », *Science* 331 (14 janvier 2011), pp. 176-182 (<http://www.sciencemag.org/content/331/6014/176.abstract>). Vidéo d'une conférence sur le thème de l'article, voir Erez Lieberman Aiden et Jean-Baptiste Michel, « What We Learned from 5 Million Books », TEDx, Cambridge, MA, 2011 (http://www.ted.com/talks/what_we_learned_from_5_million_books.html).

58. Modules sans fil placés dans les véhicules et assurance – Voir Cukier, « Data, Data Everywhere ». Jack Levis d'UPS – Entretien avec Cukier, avril 2012. Données sur les économies réalisées par UPS – Institute for Operations Research and the Management Sciences (INFORMS), « UPS Wins Gartner BI Excellence Award », 2011 (<http://www.informs.org/Announcements/UPS-wins-Gartner-BI-Excellence-Award>).

59. Recherches de Pentland – Robert Lee Hotz, « The Really Smart Phone », *Wall Street Journal*, 22 avril 2011 (<http://online.wsj.com/article/SB10001424052748704547604576263261679848814.html>).

60. Étude d'Eagle sur les bidonvilles – Nathan Eagle, « Big Data, Global Development, and Complex Systems », Santa Fe Institute, 5 mai 2010 (<http://www.youtube.com/watch?v=yaivtqlu7iM>). Également entretien avec Cukier, octobre 2012.

61. Données de Facebook – Extrait du Prospectus d'introduction en Bourse de Facebook, 2012. Données de Twitter – Alexia Tsotsis, « Twitter Is at 250 Million Tweets per Day, iOS 5 Integration Made Signups Increase 3x », *TechCrunch*, 17 octobre 2011, <http://techcrunch.com/2011/10/17/twitter-is-at-250-million-tweets-per-day/>. Utilisation de Twitter par des fonds de couverture – Kenneth Cukier, « Tracking Social Media : The Mood of the Market », *Economist.com*, 28 juin 2012 (<http://www.economist.com/blogs/graphicdetail/2012/06/tracking-social-media>).

62. Twitter et les prévisions sur les recettes de films hollywoodiens – Sitaram Asur et Bernardo A. Huberman, « Predicting the Future with Social Media », *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 492-499 ; informations en ligne <http://www.hpl.hp.com/research/scl/papers/socialmedia/socialmedia.pdf>.

Twitter et les humeurs des gens à travers le monde – Scott A. Golder et Michael W. Macy, « Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures », *Science* 333 (30 septembre 2011), pp. 1878-81.

Twitter et la vaccination contre la grippe – Marcel Salathé and Shashank Khandelwal, « Assessing Vaccination Sentiments with Online Social Media : Implications for Infectious Disease Dynamics and Control », *PLoS Computational Biology*, octobre 2011.

63. Brevet d'IBM portant sur un « sol intelligent » – Lydia Mai Do, Travis M. Grigsby, Pamela Ann Nesbitt et Lisa Anne Seacat, « Securing premises using surfaced-based computing technology », Numéro de brevet américain : 8138882. Date d'émission : 20 mars 2012.

Le mouvement « Quantified self » – « Counting Every Moment », *The Economist*, 3 mars 2012. Oreillettes conçues par Apple pour prendre des biomesures – Jesse Lee Dorogusker, Anthony Fadell, Donald J. Novotney et Nicholas R. Kalayjian, « Integrated Sensorsfor Tracking Performance Metrics », Demande de brevet américain 20090287067. Cessionnaire : Apple. Date de la demande : 23 juillet 2009. Date de publication : 11 novembre 2009.

Derawi Biometrics, « Your Walk Is Your PIN-Code », communiqué de presse du 21 février 2011 (<http://biometrics.derawi.com/?p=175>).

Information sur iTrem – Voir la page du Landmarc Research Center relative au projet iTrem sur le site du Georgia Tech Research Institute (<http://eosl.gtri.gatech.edu/Capabilities/LandmarcResearchCenter/LandmarcProjects/iTrem/tabid/798/Default.aspx>) et l'échange de courriels. Chercheurs de Kyoto sur les accéléromètres tridimensionnels – Article du site Web iMedicalApps, « GaitAnalysis Accuracy : Android App Comparable to Standard AccelerometerMethodology », *mHealth*, 23 mars 2012.

64. Les journaux ont permis la naissance de l'État-nation – Benedict Anderson, *Imagined Communities : Reflections on the Origin and Spread of Nationalism* (Verso, 2006). Les physiciens suggèrent que les informations sont à la base de tout ce qui existe – Hans Christian von Baeyer, *Information : The New Language of Science* (Harvard University Press, 2005).

Chapitre 6. Valeur

65. Histoire de Luis von Ahn – Tirée des entretiens de von Ahn avec Cukier à partir de 2010. Voir aussi Clive Thompson, « For Certain Tasks, the Cortex Still Beats the CPU », *Wired*, 25 juin 2007 (http://www.wired.com/techbiz/it/magazine/15-07/ff_humancomp?currentPage=all) ; Jessie Scanlon, « Luis vonAhn : The Pioneer of “Human Computation” », *Businessweek*, 3 novembre 2008 (<http://www.businessweek.com/stories/2008-11-03/luis-von-ahn-the-pioneer-of-human-computation-businessweek-business-news-stock-market-and-financial-advice>). Sa description technique des reCaptchas figure dans Luis von Ahn *et al.*, « reCAPTCHA : Human-Based Character Recognition via Web Security Measures », *Science* 321 (12 septembre 2008), pp. 1465-68 (<http://www.sciencemag.org/content/321/5895/1465.abstract>).

66. Exemple de la manufacture d'épingles cité par Smith – Adam Smith, *The Wealth of Nations* (réimpression, BantamClassics, 2003), livre I, chapitre un. (Une version électronique gratuite est disponible sur <http://www2.hn.psu.edu/faculty/jmanis/adam-smith/Wealth-Nations.pdf>).

67. Stockage – Viktor Mayer-Schönberger, *Delete : The Virtue of Forgetting in the Digital Age* (Princeton University Press, 2011), p. 63.

68. Utilisation de l'électricité par les voitures électriques – IBM, « IBM, Honda, and PG&E Enable Smarter Charging for Electric Vehicles », communiqué de presse, 12 avril 2012 (<http://www-03.ibm.com/press/us/en/pressrelease/37398.wss>). Voir aussi Clay Luthy, « Guest Perspective : IBM Working with PG&E to Maximize the EV Potential » *PGE Currents Magazine*, 13 avril 2012 (<http://www.pgecurrents.com/2012/04/13/ibm-working-with-pge-to-maximize-the-ev-potential>).

69. Amazon et les données d'AOL – Entretien d'Andreas Weigend avec Cukier, 2010 et 2012. Logiciel de Nuance et Google – Cukier, « Data, Data Everywhere ».

70. Entreprise de logistique – Brad Brown, Michael Chui et James Manyika, « Are You Ready for the Era of “Big Data” ? », *McKinsey Quarterly*, octobre 2011, p. 10.

71. Telefonica vend les données provenant des téléphones portables de ses abonnés – « Telefonica Hopes “Big Data” Arm Will Revive Fortunes », *BBC Online*, 9 octobre 2012. (<http://www.bbc.co.uk/news/technology-19882647>). Étude de la Danish Cancer Society – Patrizia Frei *et al.*, « Use of Mobile Phones and Risk of Brain Tumours : Update of Danish Cohort Study », *BMJ* 343 (2011) (<http://www.bmj.com/content/343/bmj.d6387>), et entretien avec Cukier, octobre 2012.

72. Les données GPS du programme Street View de Google et la voiture automate – Peter Kirwan, « This Car Drives Itself », *Wired UK*, janvier 2012 (<http://www.wired.co.uk/magazine/archive/2012/01/features/this-car-drives-itself?page=all>).

73. Correction orthographique de Google et citation – Entretien avec Cukier à Googleplex, le siège de Google à Mountain View, en Californie, décembre 2009 ; une partie de l'entretien est reprise dans Cukier, « Data, Data Everywhere ».

74. Découvertes d'Hammerbacher – Entretien avec Cukier, octobre 2012. Données issues des livres numériques de Barnes & Noble – Alexandra Alter, « Your E-Book Is Reading You », *Wall Street Journal*, 29 juin 2012 (<http://online.wsj.com/article/SB10001424052702304870304577490950051438304.html>).

75. Cours en ligne d'Andrew Ng dans le cadre de Coursera et données – Entretien avec Cukier, juin 2012.

76. Politique de libération des données publiques à l'instigation d'Obama – Barack Obama, « Memorandum présidentiel », Maison Blanche, 21 janvier 2009.

77. Valeur des données de Facebook – Pour un excellent examen de l'écart entre la valeur de marché et la valeur comptable de Facebook au moment de son introduction en Bourse, voir Doug Laney, « To Facebook You're Worth \$80.95 », *Wall Street Journal*, 3 mai 2012 (<http://blogs.wsj.com/cio/2012/05/03/to-facebook-youre-worth-80-95/>).

Pour évaluer les éléments distincts de Facebook, Laney a extrapolé à partir de la croissance de Facebook pour estimer les 2,1 billions d'éléments de contenu monétisable. Dans son article publié dans le *Wall Street Journal*, il a évalué chaque élément à trois centimes de dollar car il se fondait sur une estimation antérieure de la valorisation de Facebook correspondant à 75 milliards de dollars. Elle avait atteint plus de 100 milliards de dollars, soit cinq centimes de dollar, au moment où nous avons effectué notre propre extrapolation à partir de ses calculs. Écart de valeur entre les actifs physiques et les actifs incorporels – Steve M. Samek, « Prepared Testimony : Hearing on Adapting a 1930's Financial Reporting Model to the 21st Century », U.S. Senate Committee on Banking, Housing and Urban Affairs, Subcommittee on Securities, 19 juillet 2000. Valeur des actifs incorporels – Robert S. Kaplan et David P. Norton, *Strategy Maps : Converting Intangible Assets into Tangible Outcomes* (Harvard Business Review Press, 2004), pp. 4-5.

78. Citation de Tim O'Reilly – Entretien avec Cukier, février 2011.

Chapitre 7. Implications

79. Informations sur Decide.com – Échange de courriels entre Cukier et Etzioni, mai 2012.

80. Rapport McKinsey – James Manyika *et al.*, « Big Data : The Next Frontier for Innovation, Competition, and Productivity », McKinsey Global Institute, mai 2011 (http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation), p. 10. Citation d'Hal Varian – Entretien avec Cukier, décembre 2009.

81. Citation de Carl de Marcken – Échange de courriels avec Cukier, mai 2012.

82. MasterCard Advisors – Entretien de Gary Kearns, cadre de MasterCard Advisors, avec Cukier lors de la conférence organisée par *The Economist*, « The Ideas Economy : Information », à Santa Clara, en Californie, 8 juin 2011.

83. Accenture et la ville de St. Louis, dans le Missouri – Entretien de Cukier avec des employés municipaux, février 2007. L'Amalga Unified Intelligence System de Microsoft – « Microsoft Expands Presence in Healthcare IT Industry with Acquisition of Health Intelligence Software Azyxxi », communiqué de presse de Microsoft, 26 juillet 2006 (<http://www.microsoft.com/en-us/news/press/2006/jul06/07-26azyxxiacquisitionpr.aspx>). L'outil Amalga fait maintenant partie intégrante de Caradigm, la coentreprise créée par Microsoft avec General Electric.

84. Bradford Cross – Entretiens avec Cukier, mars-octobre 2012.

85. Amazon et le « filtrage collaboratif » – Prospectus d'introduction en Bourse, mai 1997 (<http://www.sec.gov/Archives/edgar/data/1018724/0000891020-97-000868.txt>).

86. Microprocesseurs dans les voitures – Nick Valery, « Tech. View : Cars and Software Bugs », *Economist.com*, 16 mai 2010 (http://www.economist.com/blogs/babbage/2010/05/techview_cars_and_software_bugs). Maury appelait les navires « observatoires flottants » – Maury, *The Physical Geography of the Sea*.

87. Inrix – Entretien de Cukier avec des directeurs, mai et septembre 2012.

88. Health Care Cost Institute – Sarah Kliff, « A Database That Could Revolutionize Health Care », *Washington Post*, 21 mai 2012. L'agrément d'utilisation des données établi par Decide.com – Échange de courriels entre Cukier et Etzioni, mai 2012. Marché conclu entre Google et ITA – Claire Cain Miller, « U.S. Clears Google Acquisition of Travel Software », *New York Times*, 8 avril 2011 (http://www.nytimes.com/2011/04/09/technology/09google.html?_r=0).

89. Inrix et ABS – Entretien de Cukier avec des directeurs d'Inrix, mai 2012.

90. Histoire de Roadnet et citation de Len Kennedy – Entretien avec Cukier, mai 2012. Dialogue extrait du film *Le Stratège*, réalisé par Bennett Miller, Columbia Pictures, 2011.

91. Données de McGregor correspondant à plus d'une décennie de patients-années – Entretien avec Cukier, mai 2012. Citation de Goldbloom – Entretien avec Cukier, mars 2012.

92. Box office d'Hollywood comparé aux ventes de jeux vidéo – Pour les films, voir Brooks Barnes, « A Year of Disappointment at the Movie Box Office », *New York Times*, 25 décembre 2011 (<http://www.nytimes.com/2011/12/26/business/media/a-year-of-disappointment-for-hollywood.html>). Pour les jeux vidéo, « Factbox : A Look at the \$65 billion Video Games Industry », Reuters, 6 juin 2011 (<http://uk.reuters.com/article/2011/06/06/us-videogames-factbox-idUKTRE75552I20110606>). Analyse de données chez Zynga – Nick Wingfield, « Virtual Products, Real Profits : Players Spend on Zynga's Games, but Quality Turns Some Off », *Wall Street Journal*, 9 septembre 2011 (<http://online.wsj.com/article/SB10001424053111904823804576502442835413446.html>).

93. Citation de Ken Rudin – Extraite de l'entretien de Rudin avec Niko Waesche, cité dans Erik Schlie, Jörg Rheinboldt et Niko Waesche, *Simply Seven : Seven Waysto Create a Sustainable Internet Business* (Palgrave Macmillan, 2011), p. 7. Citation de W. H. Auden – W. H. Auden, « For the Time Being », 1944. Citation de Thomas Davenport – Entretien de Davenport avec Cukier, décembre 2009. The-Numbers.com – Entretiens de Bruce Nash avec Cukier, octobre 2011 et juillet 2012.

94. Étude de Brynjolfsson – Erik Brynjolfsson, Lorin Hitt et Heekyung Kim, « Strength in Numbers : How Does Data-Driven Decisionmaking Affect Firm Performance ? », document de travail, avril 2011 (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486).

95. Rolls-Royce – Voir « Rolls-Royce : Britain's Lonely High-Flier », *The Economist*, 8 janvier 2009 (<http://www.economist.com/node/12887368>). Chiffres mis à jour communiqués par le service de presse, novembre 2012. Erik Brynjolfsson, Andrew McAfee, Michael Sorell et Feng Zhu, « Scale Without Mass : Business Process Replication and Industry Dynamics », document de travail de l'Harvard Business School, septembre 2006

(<http://www.hbs.edu/research/pdf/07-016.pdf>,
<http://hbswk.hbs.edu/item/5532.html>).

également

96. Tendence à l'augmentation du volume de données chez les gros détenteurs de données. – Voir aussi Yannis Bakos and Erik Brynjolfsson, « Bundling Information Goods : Pricing, Profits, and Efficiency », *Management Science* 45 (décembre 1999), pp. 1613-30.

97. Philip Evans – Entretiens avec les auteurs, 2011 et 2012.

Chapitre 8. Risques

98. Stasi – La plus grande partie des livres sur ce sujet est malheureusement publiée en allemand, sauf l'ouvrage bien documenté de Kristie Macrakis, *Seduced by Secrets : Inside the Stasi's Spy-Tech World* (Cambridge University Press, 2008) ; une histoire très personnelle est racontée dans Timothy Garton Ash, *The File* (Atlantic Books, 2008). Nous recommandons également le film primé aux Oscars *La Vie des autres*, réalisé par Florian Henckel von Donnersmark, Buena Vista/Sony Pictures, 2006. Caméras de surveillance près de la maison d'Orwell – « George Orwell, Big Brother Is Watching Your House », *The Evening Standard*, 31 mars 2007 (<http://www.thisislondon.co.uk/news/george-orwell-big-brother-is-watching-your-house-7086271.html>). Equifax et Experian – Daniel J. Solove, *The Digital Person : Technology and Privacy in the Information Age* (NYU Press, 2004), pp. 20-21.

99. Adresses de pâtés de maisons où résidaient des Japonais à Washington communiquées aux autorités américaines – J. R. Minkel, « The U.S. Census Bureau Gave Up Names of Japanese-Americans in WW II », *Scientific American*, 30 mars 2007 (<http://www.scientificamerican.com/article.cfm?id=confirmed-the-us-census-b>). Données utilisées par les nazis aux Pays-Bas – William Seltzer et Margo Anderson, « The Dark Side of Numbers : The Role of Population Data Systems in Human Rights Abuses », *Social Research* 68 (2001), pp. 481-513.

100. IBM et l'Holocauste – Edwin Black, *IBM et l'Holocauste. L'alliance stratégique entre l'Allemagne nazie et la plus puissante multinationale américaine* (Robert Laffont, Paris, 2001). Volume de données collectées par les compteurs d'électricité « intelligents » – Voir Elias Leake Quinn, « Smart Metering and Privacy : Existing Law and

Competing Policies ; A Report for the Colorado Public Utility Commission », printemps 2009 (http://www.w4ar.com/Danger_of_Smart_Meters_Colorado_Report.pdf).

Voir aussi Joel M. Margolis, « When Smart Grids Grow Smart Enough to Solve Crimes », Neustar, 18 mars 2010 (http://energy.gov/sites/prod/files/gcprod/documents/Neustar_Comments_DataExhibitA.pdf).

101. Fred Cate au sujet de la notification et du consentement – Fred H. Cate, « The Failure of Fair Information Practice Principles » in Jane K. Winn, éd., *Consumer Protection in the Age of the « Information Economy »* (Ashgate, 2006), p. 341 sq.

102. Diffusion de données par AOL – Michael Barbaro et Tom Zeller Jr., « A Face Is Exposed for AOL Searcher No. 4417749 », *New York Times*, 9 août 2006. Voir aussi Matthew Karnitschnig et Mylene Mangalindan, « AOL Fires Technology Chief After Web-Search Data Scandal », *Wall Street Journal*, 21 août 2006.

103. Réidentification d'une utilisatrice de Netflix – Ryan Singel, « Netflix Spilled Your *Brokeback Mountain* Secret, Lawsuit Claims », *Wired*, 17 décembre 2009 (<http://www.wired.com/threatlevel/2009/12/netflix-privacy-lawsuit>).

Diffusion de données par Netflix – Arvind Narayanan et Vitaly Shmatikov, « Robust De-Anonymization of Large Sparse Datasets », *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, p. 111 sq. (http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf) ; Arvind Narayanan et Vitaly Shmatikov, « How to Break the Anonymity of the Netflix Prize Dataset », 18 octobre 2006, arXiv : cs/0610105 [cs.CR] (<http://arxiv.org/abs/cs/0610105>).

Identification de personnes à partir de trois caractéristiques – Philippe Golle, « Revisiting the Uniqueness of Simple Demographics in the US Population », *Association for Computing Machinery Workshop on Privacy in Electronic Society 5* (2006), p. 77.

Faiblesse structurelle de l'anonymisation – Paul Ohm, « Broken Promises of Privacy : Responding to the Surprising Failure of Anonymization », *57 UCLA Law Review* 1701 (2010). Anonymité du graphe social – Lars Backstrom, Cynthia Dwork et Jon Kleinberg, « Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography », *Communications of the Association of Computing Machinery*, décembre 2011, p. 133.

104. « Boîtes noires » des automobiles – « Vehicle Data Recorders : Watching Your Driving », *The Economist*, 23 juin 2012 (<http://www.economist.com/node/21557309>). Collecte de données de la NSA – Dana Priest et William Arkin, « A Hidden World, Growing Beyond Control », *Washington Post*, 19 juillet 2010 (<http://projects.washingtonpost.com/top-secret-america/articles/a-hidden-world-growing-beyond-control/print>). Juan Gonzalez, « Whistleblower : The NSA Is Lying – U.S. Government Has Copies of Most of Your Emails », *Democracy Now*, 20 avril 2012 (http://www.democracynow.org/2012/4/20/whistleblower_the_nsa_is_lying_us). William Binney, « Sworn Declaration in the Case of Jewel v. NSA », 2 juillet 2012 (<http://publicintelligence.net/binney-nsa-declaration>). Évolution des méthodes de surveillance avec les big data – Patrick Radden Keefe, « Can Network Theory Thwart Terrorists ? », *New York Times*, 12 mars 2006 (http://www.nytimes.com/2006/03/12/magazine/312wwln_essay.html).

105. Dialogue extrait du film *Minority Report*, réalisé par Steven Spielberg, DreamWorks/20th Century Fox, 2002. Le dialogue que nous citons est à peine abrégé. Le film est fondé sur une nouvelle de Philip K. Dick publiée en 1958, mais il existe des différences importantes entre les deux versions. En particulier, la scène d'ouverture avec le mari trompé ne figure pas dans le livre, et le « casse-tête » philosophique dû à la prédiction de crimes qui n'ont pas encore été commis est présenté plus clairement dans le film de Spielberg que dans le roman. C'est pourquoi nous avons préféré établir des parallèles avec le film. Exemples de police préventive – James Vlahos, « The Department Of Pre-Crime », *Scientific American* 306 (janvier 2012), pp. 62-67.

106. Future Attribute Screening Technology (FAST) – Voir Sharon Weinberger, « Terrorist “Pre-crime” Detector Field Tested in United States », *Nature*, 27 mai 2011 (<http://www.nature.com/news/2011/110527/full/news.2011.323.html>) ; Sharon Weinberger, « Intent to Deceive », *Nature* 465 (mai 2010), pp. 412-415. Au sujet du problème des faux positifs, voir Alexander Furnas, « Homeland Security's “Pre-Crime” Screening Will Never Work », *The Atlantic Online*, 17 avril 2012 (<http://www.theatlantic.com/technology/archive/2012/04/homeland-securitys-pre-crime-screening-will-never-work/255971>).

107. Notes des lycéens et primes d'assurance – Tim Query, « Grade Inflation and the Good-Student Discount », *Contingencies Magazine*, American Academy of Actuaries, mai-juin 2007 (<http://www.contingencies.org/mayjun07/tradecraft.pdf>). Risques du profilage – Bernard E. Harcourt, *Against Prediction : Profiling, Policing, and Punishing in an Actuarial Age* (University of Chicago Press, 2006).

108. Recherche effectuée par Richard Berk – Richard Berk, « The Role of Race in Forecasts of Violent Crime », *Race and Social Problems* 1 (2009), pp. 231-242, et entretien par courriel avec Cukier, novembre 2012.

109. Passion des données chez McNamara – Phil Rosenzweig, « Robert S. McNamara and the Evolution of Modern Management », *Harvard Business Review*, décembre 2010 (<http://hbr.org/2010/12/robert-s-mcnamara-and-the-evolution-of-modern-management/ar/pr>).

110. Succès des « Petits Génies » durant la Seconde Guerre mondiale – John Byrne, *The Whiz Kids* (Doubleday, 1993). McNamara chez Ford – David Halberstam, *The Reckoning* (William Morrow, 1986), pp. 222-245.

111. Ouvrage de Kinnard – Douglas Kinnard, *The War Managers* (University Press of New England, 1977), pp. 71-25. Pour ce passage, nous avons pu mener un entretien par courriel avec le Dr. Kinnard, par l'intermédiaire de son assistant, et nous lui exprimons pour cela toute notre reconnaissance.

112. Citation « Nous croyons en Dieu – tous les autres, apportez-nous des données » – Fréquemment attribuée à W. Edwards Deming. Ted Kennedy et liste des passagers interdits de vol – Sara Kehaulani Goo, « Sen. Kennedy Flagged by No-Fly List », *Washington Post*, 20 août 2004, p. A01 (<http://www.washingtonpost.com/wp-dyn/articles/A17073-2004Aug19.html>).

113. Pratiques de recrutement chez Google – Voir Douglas Edwards, *I'm Feeling Lucky : The Confessions of Google Employee Number 59* (Houghton Mifflin Harcourt, 2011), p. 9. Voir aussi Steven Levy, *In the Plex* (Simon and Schuster, 2011), pp. 140-141. Fait ironique, les cofondateurs de Google voulaient recruter Steve Jobs au poste de DG (en dépit du fait qu'il n'était pas diplômé de l'université) ; Levy, p. 80. Tests sur 41 nuances de bleu – Laura M. Holson, « Putting a Bolder Face on Google », *New York Times*, 1^{er} mars 2009 (<http://www.nytimes.com/2009/03/01/business/01marissa.html>). Démission du concepteur principal de Google – Citation extraite (mais intégrale pour

une meilleure lisibilité) de Doug Bowman, « Goodbye, Google », billet publié sur son blog, 20 mars 2009 (<http://stopdesign.com/archive/2009/03/20/goodbye-google.html>).

114. Citation de Jobs – Steve Lohr, « Can Apple Find More Hits Without Its Tastemaker ? », *New York Times*, 18 janvier 2011, p. B1 (<http://www.nytimes.com/2011/01/19/technology/companies/19innovate.html>). Livre de Scott – James Scott, *Seeing Like a State : How Certain Schemes to Improve the Human Condition Have Failed* (Yale University Press, 1998). Citation de McNamara datant de 1967 – Extraite d’une allocution prononcée au Millsaps College à Jackson, Mississippi, citée dans *Harvard Business Review*, décembre 2010.

115. Regrets de McNamara – Robert S. McNamara avec Brian VanDeMark, *Avec le recul. La tragédie du Vietnam et ses leçons* (Le Seuil, 1998), pp. 48, 270.

Chapitre 9. Contrôle

116. Collection de livres de la bibliothèque de l’université de Cambridge – Marc Drogin, *Anathema ! Medieval Scribes and the History of Book Curses* (Allanheld and Schram, 1983), p. 37.

117. Responsabilité et protection de la vie privée – Le Center for Information Policy Leadership a lancé un projet étalé sur plusieurs années et portant sur l’interface de la responsabilité et de la protection de la vie privée ; voir http://www.informationpolicycentre.com/accountability-based_privacy_governance.

118. Dates d’expiration des données – Mayer-Schönberger, *Delete*. « Differential privacy » – Cynthia Dwork, « A Firm Foundation for Private Data Analysis », *Communications of the ACM*, janvier 2011, pp. 86-95. Facebook et forme de respect différentiel de la vie privée – A. Chin et A. Klinefelter, « Differential Privacy as a Response to the Reidentification Threat : The Facebook Advertiser Case Study », 90 *North Carolina Law Review* 1417 (2012) ; A. Haeberlen *et al.*, « Differential Privacy Under Fire », (<http://www.cis.upenn.edu/~ahae/papers/fuzz-sec2011.pdf>).

119. Entreprises suspectées de collusion – Des tentatives sont déjà en cours dans ce domaine ; voir Pim Heijnen, Marco A. Haan et Adriaan R. Soetevent, « Screening for Collusion : A Spatial Statistics Approach »,

document de travail TI 2012-058/1, Tinbergen Institute, Pays-Bas, 2012 (<http://www.tinbergen.nl/discussionpapers/12058.pdf>).

120. Désignation de représentants pour la protection des données dans certaines entreprises allemandes – Viktor Mayer-Schönberger, « Beyond Privacy, Beyond Rights : Towards a “Systems” Theory of Information Governance », 98 *California Law Review* 1853 (2010).

121. Interopérabilité des données – John Palfrey et Urs Gasser, *Interop : The Promise and Perils of Highly Interconnected Systems* (Basic Books, 2012).

Chapitre 10. Demain

122. Mike Flowers et l'analyse dans la ville de New York – Tiré de son entretien avec Cukier, juillet 2012. Une bonne description existe dans Alex Howard, « Predictive data analytics is saving lives and taxpayer dollars in New York City », *O'Reilly Media*, 26 juin 2012 (<http://strata.oreilly.com/2012/06/predictive-data-analytics-big-data-nyc.html>).

123. Walmart et Pop-Tarts – Hays, « What Wal-Mart Knows About Customers' Habits ».

124. Utilisation des big data dans les bidonvilles et pour la modélisation des mouvements de réfugiés – Nathan Eagle, « Big Data, Global Development, and Complex Systems » (<http://www.youtube.com/watch?v=yaivtqlu7iM>). Perception of time – Benedict Anderson, *Imagined Communities* (Verso, 2006).

125. « Le passé est prologue » – William Shakespeare, *La Tempête*, acte II, scène 1.

126. Expériences du CERN et stockage de données – Échange de courriels entre Cukier et des chercheurs du CERN, novembre 2012. Ordinateur de bord d'Apollo 11 – David A. Mindell, *Digital Apollo : Human and Machine in Spaceflight* (MIT Press, 2008).

a. « L'efficacité déraisonnable des données ».

b. « L'efficacité déraisonnable des mathématiques en sciences naturelles ».

Bibliographie

- Alter, Alexandra, « Your E-Book Is Reading You », *Wall Street Journal*, 29 juin 2012 (<http://online.wsj.com/article/SB10001424052702304870304577490950051438304.html>).
- Anderson, Benedict, *L'Imaginaire national. Réflexion sur l'origine et l'essor du nationalisme*, La Découverte, Paris, 2006.
- Anderson, Chris, « The End of Theory », *Wired* 16, n° 7, juillet 2008 (http://www.wired.com/science/discoveries/magazine/16-07/pb_theory).
- Asur, Sitaram et Bernardo A. Huberman, « Predicting the Future with Social Media », *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 492-499. (Une version en ligne est disponible sur <http://www.hpl.hp.com/research/scl/papers/socialmedia/socialmedia.pdf>).
- Ayres, Ian, *Super Crunchers : Why Thinking-By-Numbers Is the New Way to Be Smart*, Bantam Dell, 2007.
- Babbie, Earl, *Practice of Social Research*, 12^e éd., 2010.
- Backstrom, Lars, Cynthia Dwork et Jon Kleinberg, « Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography », *Communications of the ACM*, décembre 2011, pp. 133-141.
- Bakos, Yannis et Erik Brynjolfsson, « Bundling Information Goods : Pricing, Profits, and Efficiency », *Management Science* 45 (décembre 1999), pp. 1613-1630.
- Banko, Michele et Eric Brill, « Scaling to Very Very Large Corpora for Natural Language Disambiguation », *Microsoft Research*, 2001, p. 3 (<http://acl.ldc.upenn.edu/P/P01/P01-1005.pdf>).

- Barbaro, Michael et Tom Zeller Jr, « A Face Is Exposed for AOL Searcher No. 4417749 », *New York Times*, 9 août 2006 (<http://www.nytimes.com/2006/08/09/technology/09aol.html>).
- Barnes, Brooks, « A Year of Disappointment at the Movie Box Office », *New York Times*, 25 décembre 2011 (<http://www.nytimes.com/2011/12/26/business/media/a-year-of-disappointment-for-hollywood.html>).
- Beaty, Janice, *Seeker of Seaways : A Life of Matthew Fontaine Maury, Pioneer Oceanographer*, Pantheon Books, 1966.
- Berger, Adam L., *et al.* « The Candide System for Machine Translation », *Proceedings of the 1994 ARPA Workshop on Human Language Technology* (1994) (<http://aclweb.org/anthology-new/H/H94/H94-1100.pdf>).
- Berk, Richard, « The Role of Race in Forecasts of Violent Crime », *Race and Social Problems* 1 (2009), pp. 231-242.
- Black, Edwin, *IBM et l'Holocauste. L'alliance stratégique entre l'Allemagne nazie et la plus puissante multinationale américaine*, Robert Laffont, 2001.
- Boyd, Danah et Kate Crawford, « Six Provocations for Big Data », document de recherche présenté à l'Oxford Internet Institute's, « A Decade in Internet Time : Symposium on the Dynamics of the Internet and Society », 21 septembre 2011 (<http://ssrn.com/abstract=1926431>).
- Brown, Brad, Michael Chui et James Manyika, « Are You Ready for the Era of "Big Data" ? », *McKinsey Quarterly*, octobre 2011, p. 10.
- Brynjolfsson, Erik, Andrew McAfee, Michael Sorell et Feng Zhu, « Scale Without Mass : Business Process Replication and Industry Dynamics », HBS working paper, septembre 2006 (<http://www.hbs.edu/research/pdf/07-016.pdf> ; voir aussi <http://hbswk.hbs.edu/item/5532.html>).
- Brynjolfsson, Erik, Lorin Hitt et Heekyung Kim, « Strength in Numbers : How Does Data-Driven Decisionmaking Affect Firm Performance ? », *ICIS 2011 Proceedings*, Paper 13 (<http://aisel.aisnet.org/icis2011/proceedings/economicvalueIS/13> ; disponible également sur http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486).
- Byrne, John, *The Whiz Kids*, Doubleday, 1993.

- Cate, Fred H, « The Failure of Fair Information Practice Principles », in *Consumer Protection in the Age of the « Information Economy »*, Jane K (ed.), Ashgate, 2006.
- Winn, Jane (ed.), *Consumer Protection in the Age of the « Information Economy »*, Ashgate, 2006, p. 341 sq.
- Chin, A. et A. Klinefelter, « Differential Privacy as a Response to the Reidentification Threat : The Facebook Advertiser Case Study », 90 *North Carolina Law Review* 1417 (2012).
- Crosby, Alfred, *The Measure of Reality : Quantification and Western Society, 1250-1600*, Cambridge University Press, 1997.
- Cukier, Kenneth, « Data, Data Everywhere », *The Economist* « Special Report », 27 février 2010, pp. 1-14.
- , « Tracking Social Media : The Mood of the Market », *Economist.com*, 28 juin 2012 (<http://www.economist.com/blogs/graphicdetail/2012/06/tracking-social-media>).
- Davenport, Thomas H., Paul Barth et Randy Bean, « How “Big Data” Is Different », *Sloan Review*, 30 juillet 2012 (<http://sloanreview.mit.edu/the-magazine/2012-fall/54104/how-big-data-is-different>).
- Di Quinzio, Melanie et Anne McCarthy, « Rabies Risk Among Travellers », *CMAJ* 178, n° 5 (2008), p. 567.
- Drogin, Marc, *Anathema ! Medieval Scribes and the History of Book Curses*, Allanheld and Schram, 1983.
- Dugas, A. F., et al., « Google Flu Trends : Correlation with Emergency Department Influenza Rates and Crowding Metrics », CID Advanced Access, 8 janvier 2012, DOI 10.1093/cid/cir883.
- Duggan, Mark et Steven D. Levitt, « Winning Isn't Everything : Corruption in Sumo Wrestling », *American Economic Review* 92 (2002), pp. 1594-1605 (<http://pricetheory.uchicago.edu/levitt/Papers/DugganLevitt2002.pdf>).
- Duhigg, Charles, *Le Pouvoir des habitudes. Changer un rien pour tout changer*, Saint-Simon, 2003.
- Duhigg, Charles, « How Companies Learn Your Secrets », *New York Times*, 16 février 2012 (<http://www.nytimes.com/2012/02/19/magazine/shoppinghabits.html>).

- Dwork, Cynthia, « A Firm Foundation for Private Data Analysis », *Communications of the ACM*, janvier 2011, pp. 86-95 (<http://dl.acm.org/citation.cfm?id=1866739.1866758>).
- Economist, The, « Rolls-Royce : Britain's Lonely High-Flier », *The Economist*, 8 janvier 2009 (<http://www.economist.com/node/12887368>).
- , « Building with Big Data : The Data Revolution Is Changing the Landscape of Business », *The Economist*, 26 mai 2011 (<http://www.economist.com/node/18741392/>).
- , « Official Statistics : Don't Lie to Me, Argentina », *The Economist*, 25 février 2012 (<http://www.economist.com/node/21548242>).
- , « Counting Every Moment », *The Economist*, 3 mars 2012 (<http://www.economist.com/node/21548493>).
- , « Vehicle Data Recorders : Watching Your Driving », *The Economist*, 23 juin 2012 (<http://www.economist.com/node/21557309>).
- Edwards, Douglas, *I'm Feeling Lucky : The Confessions of Google Employee Number 59*, Houghton Mifflin Harcourt, 2011.
- Ehrenberg, Rachel, « Predicting the Next Deadly Manhole Explosion », *Wired*, 7 juillet 2010 (<http://www.wired.com/wiredscience/2010/07/manholeexplosions>).
- Eisenstein, Elizabeth L., *La Révolution de l'imprimé dans l'Europe des premiers temps modernes*, La Découverte, 1991.
- Etzioni, Oren, C. A. Knoblock, R. Tuchinda et A. Yates, « To Buy or Not to Buy : Mining Airfare Data to Minimize Ticket Purchase Price », *SIGKDD '03*, 24-27 août 2003 (<http://knight.cis.temple.edu/~yates/papers/hamlet-kdd03.pdf>).
- Frei, Patrizia, *et al.*, « Use of Mobile Phones and Risk of Brain Tumours : Update of Danish Cohort Study », *BMJ* 2011, 343 (<http://www.bmj.com/content/343/bmj.d6387>).
- Furnas, Alexander, « Homeland Security's "Pre-Crime" Screening Will Never Work », *The Atlantic Online*, 17 avril 2012 (<http://www.theatlantic.com/technology/archive/2012/04/homeland-securitys-pre-crime-screeningwill-never-work/255971/>).
- Garton Ash, Timothy, *The File*, Atlantic Books, 2008.
- Geron, Tomio, « Twitter's Dick Costolo : Twitter Mobile Ad Revenue Beats Desktop on Some Days », *Forbes*, 6 juin 2012 (<http://www.forbes.com/sites/tomiogeron/2012/06/06/twitters-dick-costolo-mobile-ad-revenue-beats-desktop-on-some-days>).

- Ginsburg, Jeremy, *et al.*, « Detecting Influenza Epidemics Using Search Engine Query Data », *Nature* 457 (2009), pp. 1012-14 (<http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html>).
- Golder, Scott A. et Michael W. Macy, « Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures », *Science* 333 (30 septembre 2011), pp. 1878-81.
- Golle, Philippe, « Revisiting the Uniqueness of Simple Demographics in the US Population », *Association for Computing Machinery Workshop on Privacy in Electronic Society* 5 (2006), pp. 77-80.
- Goo, Sara Kehaulani, « Sen. Kennedy Flagged by No-Fly List », *Washington Post*, 20 août 2004, p. A01 (<http://www.washingtonpost.com/wp-dyn/articles/A17073-2004Aug19.html>).
- Haerberlen, A., *et al.*, « Differential Privacy Under Fire », in *SEC'11 : Proceedings of the 20th USENIX conference on Security*, p. 33 (<http://www.cis.upenn.edu/~ahae/papers/fuzz-sec2011.pdf>).
- Halberstam, David, *The Reckoning*, William Morrow, 1986.
- Haldane, J. B. S., « On Being the Right Size », *Harper's Magazine*, mars 1926 (<http://harpers.org/archive/1926/03/on-being-the-right-size/>).
- Halevy, Alon, Peter Norvig et Fernando Pereira, « The Unreasonable Effectiveness of Data », *IEEE Intelligent Systems*, mars/avril 2009, pp. 8-12.
- Harcourt, Bernard E., *Against Prediction : Profiling, Policing, and Punishing in an Actuarial Age*, University of Chicago Press, 2006.
- Hardy, Quentin, « Bizarre Insights from Big Data », *NYTimes.com*, 28 mars 2012 (<http://bits.blogs.nytimes.com/2012/03/28/bizarre-insights-from-big-data>).
- Hays, Constance L., « What Wal-Mart Knows About Customers' Habits », *New York Times*, 14 novembre 2004 (<http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html>).
- Hearn, Chester G., *Tracks in the Sea : Matthew Fontaine Maury and the Mapping of the Oceans*, International Marine/McGraw-Hill, 2002.
- Helland, Pat, « If You Have Too Much Data then “Good Enough” Is Good Enough », *Communications of the ACM*, juin 2011, p. 40 sq.
- Hilbert, Martin et Priscilla López, « The World's Technological Capacity to Store, Communicate, and Compute Information », *Science* 1

(avril 2011), pp. 60-65.

- , « How to Measure the World's Technological Capacity to Communicate, Store and Compute Information ? », *International Journal of Communication* 2012, pp. 1042-55 (ijoc.org/ojs/index.php/ijoc/article/viewFile/1562/742).
- Holson, Laura M., « Putting a Bolder Face on Google », *New York Times*, 1^{er} mars 2009, p. BU 1 (<http://www.nytimes.com/2009/03/01/business/01marissa.html>).
- Hopkins, Brian et Boris Evelson, « Expand Your Digital Horizon with Big Data », Forrester, 30 septembre 2011.
- Hotz, Robert Lee, « The Really Smart Phone », *Wall Street Journal*, 22 avril 2011 (<http://online.wsj.com/article/SB10001424052748704547604576263261679848814.html>).
- Hutchins, John, « The First Public Demonstration of Machine Translation : The Georgetown-IBM System, 7th January 1954 », novembre 2005 (<http://www.hutchinsweb.me.uk/GU-IBM-2005.pdf>).
- Inglehart, R. et H. D. Klingemann, *Genes, Culture, Democracy and Happiness*, MIT Press, 2000.
- Isaacson, Walter, *Steve Jobs*, J.-C. Lattès, Paris, 2011.
- Kahneman, Daniel, *Système 1, système 2. Les deux vitesses de la pensée*, Flammarion, Paris, 2012.
- Kaplan, Robert S. et David P. Norton, *Strategy Maps : Converting Intangible Assets into Tangible Outcomes*, Harvard Business Review Press, 2004.
- Karnitschnig, Matthew et Mylene Mangalindan, « AOL Fires Technology Chief After Web-Search Data Scandal », *Wall Street Journal*, 21 août 2006.
- Keefe, Patrick Radden, « Can Network Theory Thwart Terrorists ? », *New York Times*, 12 mars 2006 (http://www.nytimes.com/2006/03/12/magazine/312wwln_essay.html).
- Kinnard, Douglas, *The War Manager*, University Press of New England, 1977.
- Kirwan, Peter, « This Car Drives Itself », *Wired UK*, janvier 2012 (<http://www.wired.co.uk/magazine/archive/2012/01/features/this-car-drives-itself>).

- Kliff, Sarah, « A Database That Could Revolutionize Health Care », *Washington Post*, 21 mai 2012.
- Kruskal, William et Frederick Mosteller, « Representative Sampling, IV : The History of the Concept in Statistics, 1895-1939 », *International Statistical Review* 48 (1980), pp. 169-195.
- Laney, Doug, « To Facebook You're Worth \$80.95 », *Wall Street Journal*, 3 mai 2012 (<http://blogs.wsj.com/cio/2012/05/03/to-facebook-youre-worth-80-95>).
- Latour, Bruno, *Pasteur. Guerre et paix des microbes*, La Découverte, 2011.
- Levitt, Steven D. et Stephen J. Dubner, *Freakonomics*, Denoël, 2006.
- Levy, Steven, *In the Plex*, Simon and Schuster, 2011.
- Lewis, Charles Lee, *Matthew Fontaine Maury : The Pathfinder of the Seas*, U.S. Naval Institute, 1927.
- Lohr, Steve, « Can Apple Find More Hits Without Its Tastemaker ? », *New York Times*, 18 janvier 2011, p. B1 (<http://www.nytimes.com/2011/01/19/technology/companies/19innovate.html>).
- Lowrey, Annie, « Economists' Programs Are Beating U.S. at Tracking Inflation », *Washington Post*, 25 décembre 2010 (<http://www.washingtonpost.com/wp-dyn/content/article/2010/12/25/AR2010122502600.html>).
- Macrakis, Kristie, *Seduced by Secrets : Inside the Stasi's Spy-Tech World*, Cambridge University Press, 2008.
- Manyika, James, *et al.*, « Big Data : The Next Frontier for Innovation, Competition, and Productivity », *McKinsey Global Institute*, mai 2011 (http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation).
- Marcus, James, *Amazonia : Five Years at the Epicenter of the Dot.Com Juggernaut*, The New Press, 2004.
- Margolis, Joel M., « When Smart Grids Grow Smart Enough to Solve Crimes », *Neustar*, 18 mars 2010 (http://energy.gov/sites/prod/files/gcprod/documents/Neustar_Comments_DataExhibitA.pdf).
- Maury, Matthew Fontaine, *Géographie physique de la mer*, Librairie J. Corbéard, 1858.
- Mayer-Schönberger, Viktor, « Beyond Privacy, Beyond Rights : Towards a "Systems" Theory of Information Governance », 98 *California Law*

- Review 1853* (2010).
- , *Delete : The Virtue of Forgetting in the Digital Age*, Princeton University Press, 2^e éd., 2011.
- McGregor, Carolyn, Christina Catley, Andrew James et James Padbury, « Next Generation Neonatal Health Informatics with Artemis », in European Federation for Medical Informatics, *User Centred Networked Health Care*, ed. A. Moen *et al.* (IOS Press, 2011), p. 117 sq.
- McNamara, Robert S., avec Brian VanDeMark, *Avec le recul. La tragédie du Vietnam et ses leçons*, Le Seuil, 1998.
- Mehta, Abhishek, « Big Data : Powering the Next Industrial Revolution », tableau Software White Paper, 2011.
- Michel, Jean-Baptiste, *et al.* « Quantitative Analysis of Culture Using Millions of Digitized Books », *Science* 331 (14 janvier 2011), pp. 176-182 (<http://www.sciencemag.org/content/331/6014/176.abstract>).
- Miller, Claire Cain, « U.S. Clears Google Acquisition of Travel Software », *New York Times*, 8 avril 2011 (http://www.nytimes.com/2011/04/09/technology/09google.html?_r=0).
- Mills, Howard, « Analytics : Turning Data into Dollars », *Forward Focus*, décembre 2011 (http://www.deloitte.com/assets/Dcom-UnitedStates/Local%20Assets/Documents/FSI/US_FSI_Forward%20Focus_Analytics_Turning%20data%20into%20dollars_120711.pdf).
- Mindell, David A., *Digital Apollo : Human and Machine in Spaceflight*, MIT Press, 2008.
- Minkel, J. R., « The U.S. Census Bureau Gave Up Names of Japanese-Americans in WW II », *Scientific American*, 30 mars 2007 (<http://www.scientificamerican.com/article.cfm?id=confirmed-the-us-census-b>).
- Murray, Alexander, *Reason and Society in the Middle Ages*, Oxford University Press, 1978.
- Nalimov, E. V., G. McC. Haworth et E. A. Heinz, « Space-Efficient Indexing of Chess Endgame Tables », *ICGA Journal* 23, n° 3 (2000), pp. 148-162.
- Narayanan, Arvind et Vitaly Shmatikov, « How to Break the Anonymity of the Netflix Prize Dataset », 18 octobre 2006, arXiv : cs/0610105 (<http://arxiv.org/abs/cs/0610105>).
- , « Robust De-Anonymization of Large Sparse Datasets », *Proceedings of the 2008 IEEE Symposium on Security and Privacy*,

- p. 111 (http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf).
- Nazareth, Rita et Julia Leite, « Stock Trading in U.S. Falls to Lowest Level Since 2008 », *Bloomberg*, 13 août 2012 (<http://www.bloomberg.com/news/2012-08-13/stock-trading-in-u-s-hits-lowest-level-since-2008-as-vix-falls.html>).
- Negroponte, Nicholas, *L'Homme numérique*, Robert Laffont, 1995.
- Neyman, Jerzy, « On the Two Different Aspects of the Representative Method : The Method of Stratified Sampling and the Method of Purposive Selection », *Journal of the Royal Statistical Society* 97, n° 4 (1934), pp. 558-625.
- Ohm, Paul, « Broken Promises of Privacy : Responding to the Surprising Failure of Anonymization », *57 UCLA Law Review* 1701 (2010).
- Onnela, J. P., *et al.*, « Structure and Tie Strengths in Mobile Communication Networks », *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 104 (may 2007), pp. 7332-36 (<http://nd.edu/~dddas/Papers/PNAS0610245104v1.pdf>).
- Palfrey, John et Urs Gasser, *Interop : The Promise and Perils of Highly Interconnected Systems*, Basic Books, 2012.
- Pearl, Judea, *Causality : Models, Reasoning and Inference*, 2^e éd., Cambridge University Press, 2009.
- President's Council of Advisors on Science and Technology, « Report to the President and Congress, Designing a Digital Future : Federally Funded Research and Development in Networking and Information Technology », décembre 2010 (<http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf>).
- Priest, Dana et William Arkin, « A Hidden World, Growing Beyond Control », *Washington Post*, 19 juillet 2010 (<http://projects.washingtonpost.com/top-secret-america/articles/a-hidden-world-growing-beyond-control/print>).
- Query, Tim, « Grade Inflation and the Good-Student Discount », *Contingencies Magazine*, American Academy of Actuaries, mai-juin 2007 (<http://www.contingencies.org/mayjun07/tradecraft.pdf>).
- Quinn, Elias Leake, « Smart Metering and Privacy : Existing Law and Competing Policies ; A Report for the Colorado Public Utility Commission », printemps 2009

- (http://www.w4ar.com/Danger_of_Smart_Meters_Colorado_Report.pdf).
- Reshef, David, *et al.*, « Detecting Novel Associations in Large Data Sets », *Science* (2011), pp. 1518-24.
- Rosenthal, Jonathan, « Banking Special Report », *The Economist*, 19 mai 2012, pp. 7-8.
- Rosenzweig, Phil, « Robert S. McNamara and the Evolution of Modern Management », *Harvard Business Review*, décembre 2010, pp. 87-93 (<http://hbr.org/2010/12/robert-s-mcnamara-and-the-evolution-of-modern-management/ar/pr>).
- Rudin, Cynthia, *et al.*, « 21st-Century Data Miners Meet 19th-Century Electrical Cables », *Computer*, juin 2011, pp. 103-105.
- , « Machine Learning for the New York City Power Grid », *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.2 (2012), pp. 328-345 (<http://hdl.handle.net/1721.1/68634>).
- Rys, Michael, « Scalable SQL », *Communications of the ACM*, juin 2011, 48, pp. 48-53.
- Salathé, Marcel et Shashank Khandelwal, « Assessing Vaccination Sentiments with Online Social Media : Implications for Infectious Disease Dynamics and Control », *PLoS Computational Biology* 7, n° 10 (octobre 2011).
- Savage, Mike et Roger Burrows, « The Coming Crisis of Empirical Sociology », *Sociology* 41 (2007), pp. 885-899.
- Schlie, Erik, Jörg Rheinboldt et Niko Waesche, *Simply Seven : Seven Ways to Create a Sustainable Internet Business*, Palgrave Macmillan, 2011.
- Scanlon, Jessie, « Luis von Ahn : The Pioneer of “Human Computation” », *Businessweek*, 3 novembre 2008 (<http://www.businessweek.com/stories/2008-11-03/luis-von-ahn-the-pioneer-of-human-computation-businessweek-business-news-stock-market-and-financial-advice>).
- Scism, Leslie et Mark Maremont. « Inside Deloitte’s Life-Insurance Assessment Technology », *Wall Street Journal*, 19 novembre 2010 (<http://online.wsj.com/article/SB10001424052748704104104575622531084755588.html>).
- , « Insurers Test Data Profiles to Identify Risky Clients », *Wall Street Journal*, 19 novembre 2010

(<http://online.wsj.com/article/SB10001424052748704648604575620750998072986.html>).

Scott, James, *Seeing Like a State : How Certain Schemes to Improve the Human Condition Have Failed*, Yale University Press, 1998.

Seltzer, William et Margo Anderson, « The Dark Side of Numbers : The Role of Population Data Systems in Human Rights Abuses », *Social Research* 68 (2001) pp. 481-513.

Silver, Nate, *The Signal and the Noise : Why So Many Predictions Fail – But Some Don't*, Penguin, 2012.

Singel, Ryan, « Netflix Spilled Your Brokeback Mountain Secret, Lawsuit Claims », *Wired*, 17 décembre 2009 (<http://www.wired.com/threatlevel/2009/12/netflix-privacy-lawsuit>).

Smith, Adam, *Recherche sur la nature et les causes de la richesse des nations*, (2 tomes) coll. « Garnier-Flammarion », Flammarion, 1999.

Solove, Daniel J., *The Digital Person : Technology and Privacy in the Information Age*, NYU Press, 2004.

Surowiecki, James, « A Billion Prices Now », *New Yorker*, 30 mai 2011 (http://www.newyorker.com/talk/financial/2011/05/30/110530ta_talk_surowiecki).

Taleb, Nassim Nicholas, *Le Hasard sauvage. Comment la chance nous trompe*, Les Belles Lettres, 2009.

———, *Le Cygne noir. La puissance de l'imprévisible*, Les Belles Lettres, 2008.

Thompson, Clive, « For Certain Tasks, the Cortex Still Beats the CPU », *Wired*, 25 juin 2007 (http://www.wired.com/techbiz/it/magazine/15-07/ff_humancomp?currentPage=all).

Thurm, Scott, « Next Frontier in Credit Scores : Predicting Personal Behavior », *Wall Street Journal*, 27 octobre 2011 (<http://online.wsj.com/article/SB10001424052970203687504576655182086300912.html>).

Tsotsis, Alexia, « Twitter Is at 250 Million Tweets per Day, iOS 5 Integration Made Signups Increase 3x », *TechCrunch*, 17 octobre 2011 (<http://techcrunch.com/2011/10/17/twitter-is-at-250-million-tweets-per-day>).

Valery, Nick, « Tech.View : Cars and Software Bugs », *The Economist*, 16 mai 2010

- (http://www.economist.com/blogs/babbage/2010/05/techview_cars_and_software_bugs).
- Vlahos, James, « The Department Of Pre-Crime », *Scientific American* 306 (janvier 2012), pp. 62-67.
- Von Baeyer, Hans Christian, *Information : The New Language of Science*, Harvard University Press, 2005.
- Von Ahn, Luis, *et al.*, « reCAPTCHA : Human-Based Character Recognition via Web Security Measures », *Science* 321 (12 septembre 2008), pp. 1465-68 (<http://www.sciencemag.org/content/321/5895/1465.abstract>).
- Watts, Duncan, *Everything Is Obvious Once You Know the Answer : How Common Sense Fails Us*, Atlantic, 2011.
- Weinberger, David, *Everything Is Miscellaneous : The Power of the New Digital Disorder*, Times Books, 2007.
- Weinberger, Sharon, « Intent to Deceive », *Nature* 465 (mai 2010), pp. 412-415 (<http://www.nature.com/news/2010/100526/full/465412a.html>).
- , « Terrorist “Pre-crime” Detector Field Tested in United States », *Nature*, 27 mai 2011 (<http://www.nature.com/news/2011/110527/full/news.2011.323.html>).
- Whitehouse, David, « UK Science Shows Cave Art Developed Early », *BBC News Online*, 3 octobre 2011 (<http://news.bbc.co.uk/1/hi/sci/tech/1577421.stm>).
- Wigner, Eugene, « The Unreasonable Effectiveness of Mathematics in the Natural Sciences », *Communications on Pure and Applied Mathematics* 13, n° 1 (1960), pp. 1-14.
- Wilks, Yorick, *Machine Translation : Its Scope and Limits*, Springer, 2008.
- Wingfield, Nick, « Virtual Products, Real Profits : Players Spend on Zynga’s Games, but Quality Turns Some Off », *Wall Street Journal*, 9 septembre 2011 (<http://online.wsj.com/article/SB10001424053111904823804576502442835413446.html>).

Remerciements

Nous avons eu tous deux la chance de collaborer avec Lewis M. Branscomb et de beaucoup apprendre auprès de cet homme qui a été très tôt un grand dans le domaine des réseaux d'information et de l'innovation. Son intelligence, son éloquence, son énergie, son professionnalisme, son humour et son inépuisable curiosité demeurent une source d'inspiration pour nous. Quant à sa partenaire si sympathique et si perspicace, Connie Mullin, nous lui présentons nos excuses pour ne pas avoir suivi sa suggestion d'intituler le livre *Superdonnées*.

Momin Malik, grâce à ses exceptionnelles qualités d'intelligence et d'assiduité au travail, a été un excellent assistant de recherche. Nous avons le privilège d'être représentés par Lisa Adams et David Miller de la Garamond Agency, qui ont été tout simplement magnifiques sur tous les plans. Eamon Dolan, notre éditeur, a été fantastique. Digne représentant de cette espèce rare d'éditeurs qui sait à la perfection réviser le texte et stimuler la réflexion, il nous a permis d'aboutir à des résultats bien meilleurs que ce que nous espérions. Nous remercions toute l'équipe de la maison d'édition Houghton Mifflin Harcourt, et en particulier Beth Burleigh Fuller et Ben Hyman. Nos remerciements vont également à Camille Smith, notre experte relectrice. Nous sommes reconnaissants à James Fransham, de l'hebdomadaire *The Economist* pour son excellente vérification des faits relatés dans le manuscrit et pour ses critiques avisées.

Nous sommes tout particulièrement reconnaissants à tous ces praticiens des big data qui nous ont consacré du temps pour nous expliquer leur travail, notamment Oren Etzioni, Cynthia Rudin, Carolyn McGregor et Mike Flowers.

Viktor a des remerciements personnels à adresser :

Je remercie Philip Evans qui, dans le dialogue que nous entretenons depuis plus de dix ans, a toujours eu une longueur d'avance sur moi dans sa réflexion et qui a montré une précision et une éloquence remarquables dans l'expression de ses idées.

Je suis également reconnaissant à mon ancien collègue David Lazer, universitaire compétent et très tôt spécialisé en big data. Je l'ai consulté trente-six mille fois.

Je remercie tous les participants de l'Oxford Digital Data Dialogue de 2011 (rencontre sur le thème des big data), et plus particulièrement son coprésident Fred Cate, avec qui j'ai eu des discussions extrêmement fructueuses.

Grâce à la présence de bon nombre de chercheurs dans le domaine des big data parmi mes collègues, l'Oxford Internet Institute, au sein duquel je travaille, a offert un environnement particulièrement adéquat pour la rédaction de ce livre. Je n'aurais pu imaginer de meilleur lieu pour l'écrire. J'exprime aussi ma gratitude envers le Keble College, où je suis enseignant-chercheur, pour son appui sans lequel je n'aurais pas eu accès à une partie des sources essentielles utilisées.

Lorsque quelqu'un écrit un livre, c'est sa famille qui paie le plus lourd tribut. Outre les nombreuses heures passées dans mon bureau devant l'écran de mon ordinateur, il y a eu aussi celles encore plus nombreuses où je restais plongé dans mes pensées tout en étant physiquement présent. Et pour cela, je dois demander pardon à mon épouse Birgit et mon petit Viktor. Je promets de faire plus d'efforts à l'avenir.

Kenn souhaite lui aussi exprimer ses remerciements :

Je suis reconnaissant aux nombreux scientifiques spécialistes des données qui m'ont apporté leur contribution, en particulier Jeff Hammerbacher, Amr Awadallah, DJ Patil, Michael Driscoll, Michael Freed, ainsi qu'à d'autres nombreuses personnes de l'équipe de Google au fil des ans (je citerai notamment Hal Varian, Jeremy Ginsberg, Peter Norvig et Udi Manber, mais aussi Eric Schmidt et Larry Page avec qui j'ai eu quelques chats très précieux mais hélas trop brefs).

Ma réflexion s'est enrichie grâce à Tim O'Reilly, une autorité de l'ère d'Internet, et Marc Benioff, de Salesforce.com, qui a été un véritable enseignant pour moi. Les éclairages apportés par Matthew Hindman ont été inestimables, comme d'habitude. L'aide de James Guszczka de Deloitte a été

incroyablement précieuse, tout comme celle de Geoff Hyatt, ami de longue date et créateur d'entreprises en série dans le domaine des données. J'adresse des remerciements tout particuliers à Pete Warden, à la fois philosophe et praticien des big data.

De nombreux amis m'ont proposé leurs idées et prodigué leurs conseils, notamment John Turner, Angelika Wolf, Niko Waesche, Katia Verresen, David Wishart, Anna Petherick, Blaine Harden et Jessica Kowal. D'autres m'ont inspiré certains thèmes du livre ; je citerai Blaise Aguera y Arcas, Eric Horvitz, David Auerbach, Gil Elbaz, Tyler Bell, Andrew Wyckoff tout comme de nombreux membres à l'OCDE, Stephen Brobst et l'équipe de Teradata, Anthony Goldbloom et Jeremy Howard de Kaggle, Edd Dumbill, Roger Magoulas et l'équipe d'O'Reilly Media, et enfin Edward Lazowska. James Cortada est digne du Panthéon ! Merci également à Ping Li d'Accel Partners et Roger Ehrenberg d'IA Ventures.

Au sein de l'hebdomadaire *The Economist*, mes collègues m'ont fait part d'idées géniales et m'ont apporté un formidable soutien. Je remercie en particulier mes rédacteurs en chef Tom Standage, Daniel Franklin, et John Micklethwait, ainsi que Barbara Beck qui a dirigé le dossier spécial « Des données, des données de toutes parts » à l'origine de ce livre. Henry Tricks et Dominic Zeigler, mes collègues à Tokyo, ont été pour moi des modèles à suivre dans le sens où ils sont toujours en quête de nouveauté et qu'ils savent admirablement les présenter. Comme de coutume, Oliver Morton m'a fait bénéficier de suggestions avisées dans les moments où j'en avais le plus besoin.

L'organisation autrichienne Salzburg Global Seminar m'a offert une merveilleuse opportunité de questionnement intellectuel et des conditions de tranquillité idéales pour écrire et réfléchir. Une table ronde organisée en juillet 2011 par l'Aspen Institut a permis de faire émerger de nombreuses idées ; j'en remercie les participants et l'organisateur, Charlie Firestone. Ma reconnaissance va également à Teri Elniski pour son formidable soutien.

Frances Cairncross, recteur du collège d'Exeter, à Oxford, m'a beaucoup encouragé et a mis à ma disposition un lieu tranquille où séjourner. J'ai concentré mon esprit sur des questions de technologie et de société en me fondant en toute modestie sur celles qu'elle avait traitées dans son ouvrage *The Death of Distance*, quinze ans auparavant, qui avait inspiré le jeune journaliste que j'étais à l'époque. Il a été gratifiant de traverser chaque matin la cour du Collège d'Exeter en sachant que j'avais

probablement repris le flambeau qu'elle avait porté, même si dans sa main, il avait brûlé avec bien plus d'éclat.

Je tiens à exprimer ma profonde gratitude à ma famille qui a dû me supporter – ou le plus souvent supporter mon absence. Mes parents, ma sœur et d'autres membres de la famille méritent certes des remerciements mais je réserve la plus grande part de ma reconnaissance à mon épouse Heather et nos enfants Charlotte et Kaz, car il ne m'aurait pas été possible d'écrire ce livre sans leur soutien, leurs encouragements et leurs idées.

Nous sommes tous deux reconnaissants à ces nombreuses personnes avec qui nous avons eu des discussions sur le thème des big data, bien avant que le terme lui-même n'ait été vulgarisé. À ce propos, nous remercions tout particulièrement les participants des éditions successives de la Rueschlikon Conference on Information Policy, dont Viktor a été le coorganisateur et Kenn le rapporteur. Nous remercions encore et en particulier Joseph Alhadeff, Bernard Benhamou, John Seely Brown, Herbert Burkert (qui nous ont fait découvrir le commandant Maury), Peter Cullen, Ed Felten, Urs Gasser, Joi Ito, Jeff Jonas, Nicklas Lundblad, Douglas Merrill, Rick Murray, Cory Ondrejka et Paul Schwartz.

Viktor Mayer-Schönberger
Kenneth Cukier
Oxford/Londres, août 2012

Index

A

abaques [1](#), [2](#)

Accenture [1](#), [2](#), [3](#)

activités en ligne. *Voir* commerce électronique

Acxiom [1](#), [2](#)

AirSage [1](#), [2](#)

algorithmes, ordinateur [1](#), [2](#), [3](#), [4](#)

améliorations dans les [1](#), [2](#), [3](#), [4](#)

transparence des [1](#)

« algorithmistes » [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#)

Allemagne de l'Est : en tant qu'État policier [1](#), [2](#), [3](#), [4](#)

Alta Vista [1](#)

Amazon [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)

comptes-rendus de livres chez [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)

en tant que société des big data [1](#), [2](#)

et livres numériques [1](#), [2](#), [3](#), [4](#)

filtrage collaboratif chez [1](#), [2](#), [3](#)

réutilisation de données par [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

Amazonia (Marcus) [1](#)

analyse de données. *Voir aussi* analyse des corrélations \; analyse prédictive

McNamara et l' [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#)

par opposition à intuition [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#)

sociétés axées sur l' [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)

analyse des conditions de circulation : par Inrix [1](#), [2](#), [3](#)

analyse des corrélations. *Voir aussi* analyse de données \; analyse prédictive

dans des dossiers médicaux [1](#), [2](#), [3](#), [4](#)

dans du texte [1](#), [2](#)

dans la conception de jeux vidéo [1](#), [2](#)

dans la navigation maritime [1](#), [2](#)

dans les données sur les ventes [1](#)

des informations [1](#), [2](#), [3](#)

données substitutives dans l' [1](#), [2](#), [3](#), [4](#), [5](#)
et « fin de la théorie » [1](#), [2](#), [3](#)
et big data [1](#), [2](#), [3](#), [4](#)
et pointages de crédit [1](#)
et scandale des « subprimes » (2009) [1](#)
fondée sur l'hypothèse [1](#), [2](#), [3](#), [4](#)
non linéaires [1](#), [2](#)
par opposition à méthode scientifique [1](#), [2](#)

analyse des réseaux [1](#)

analyse des réseaux sociaux : Huberman et [1](#)

analyse prédictive. *Voir aussi* analyse des corrélations \; analyse de données
[1](#), [2](#)

big data et [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#)
châtiment basé sur l' [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#)
commissions de libération conditionnelle utilisent l' [1](#)
dans le domaine de la santé [1](#), [2](#), [3](#), [4](#)
dans le domaine des sports [1](#), [2](#), [3](#), [4](#), [5](#)
dans le profilage [1](#)
dans le secteur des assurances [1](#), [2](#)
dans les défaillances mécaniques et structurelles [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)
Département de la sécurité intérieure des États-Unis utilise l' [1](#)
durant la guerre d'Irak [1](#)
et terrorisme [1](#), [2](#), [3](#), [4](#)
par opposition à libre arbitre [1](#), [2](#), [3](#), [4](#), [5](#)
par Target [1](#), [2](#)
par UPS [1](#)
police utilise l' [1](#), [2](#), [3](#), [4](#)

Anderson, Chris : au sujet de la « fin de la théorie » [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

anonymisation : les big data mettent en échec l' [1](#), [2](#), [3](#)

de données [1](#), [2](#)
protection de la vie privée et [1](#), [2](#)

Antiquité : comptage dans l' [1](#), [2](#)

AOL : n'a pas compris l'importance de la réutilisation de données [1](#)

a publié des données personnelles [1](#), [2](#), [3](#)

Apple [1](#)

et données récupérées de téléphones portables [1](#), [2](#), [3](#)

applications de données : sociétés axées sur les innovations dans les [1](#), [2](#), [3](#),
[4](#), [5](#), [6](#), [7](#), [8](#), [9](#)

Apprentissage machine [1](#), [2](#)

Arnold, Thelma [1](#)

Assurances : analyse prédictive dans le secteur des [1](#), [2](#)

utilisent des données de géolocalisation [1](#), [2](#), [3](#)

Asthmapolis [1](#)

astronomie : big data en [1](#)

automobiles, automates [1](#), [2](#), [3](#), [4](#)

Avec le recul (McNamara) [1](#)

Aviva [1](#), [2](#)

Ayres, Ian : *Super Crunchers* [1](#)

B

Bacon, Francis [1](#)

Banko, Michele [1](#), [2](#), [3](#), [4](#), [5](#)

Banque mondiale [1](#)
et données ouvertes [1](#)

Barabási, Albert-László [1](#), [2](#)

Barnes & Noble [1](#), [2](#)

Basis [1](#)

Beane, Billy [1](#), [2](#)

Bell Labs [1](#)

Berners-Lee, Tim [1](#)

Bezos, Jeff [1](#), [2](#), [3](#), [4](#)

big data. *Voir aussi* données \; informations \; données ouvertes

« côté sombre » des [1](#), [2](#), [3](#), [4](#)

« tournure d'esprit ». *Voir* applications des données

analyse de corrélations et [1](#), [2](#), [3](#), [4](#)

chaîne de valeur [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#)

collecte et gestion de [1](#), [2](#)

comme source d'avantage concurrentiel [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

dans la théorie des réseaux [1](#), [2](#)

dans le commerce électronique [1](#), [2](#), [3](#), [4](#)

dans le développement économique [1](#), [2](#)

dans le domaine de la finance [1](#), [2](#), [3](#), [4](#)

dans le domaine de la santé [1](#), [2](#), [3](#), [4](#)

dans le séquençage de l'ADN [1](#)

dans les prêts personnels [1](#), [2](#)

effets psychologiques des [1](#), [2](#), [3](#)

effets sociaux et économiques des [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18](#), [19](#), [20](#), [21](#), [22](#)

en astronomie [1](#)

et « explicabilité » [1](#)

et analyse prédictive [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)

et calcul de l'inflation [1](#), [2](#)

et causalité [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)

et changements climatiques [1](#)

et création de valeur économique [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#)

et fraude à la carte de crédit [1](#), [2](#), [3](#), [4](#)

et intelligence artificielle [1](#)

et modèles d'entreprise [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)
et raffineries de pétrole [1](#)
et réglementation antitrust [1](#), [2](#)
et réglementation gouvernementale [1](#), [2](#), [3](#), [4](#), [5](#)
et responsabilité individuelle [1](#), [2](#), [3](#)
et voitures électriques [1](#), [2](#)
éthique en matière de [1](#), [2](#)
exactitude et [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)
imprécision en tant que caractère positif des [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#)
lois contre l'utilisation abusive des [1](#), [2](#), [3](#)
mettent en échec l'anonymisation [1](#), [2](#), [3](#), [4](#), [5](#)
modifient les perceptions humaines [1](#)
nature des [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#)
ont une base théorique [1](#), [2](#), [3](#)
protection de la vie privée et [1](#), [2](#), [3](#), [4](#), [5](#)
remplacent l'échantillonnage statistique [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)
rôle de l'expertise en la matière dans les [1](#), [2](#), [3](#), [4](#)
utilisation abusive potentielle des [1](#), [2](#), [3](#), [4](#), [5](#)

Billion Prices Project [1](#), [2](#)

Bing [1](#)

Binney, William [1](#)

Bloomberg, Michael [1](#), [2](#)

Boston Consulting Group [1](#)

Bowman, Douglas [1](#), [2](#)

Boyd, Danah [1](#)

Brahe, Tycho [1](#)

Brill, Eric [1](#), [2](#), [3](#), [4](#), [5](#)

Brin, Sergey [1](#)

British Petroleum [1](#), [2](#)

Brynjolfsson, Erik [1](#), [2](#)

Bureau américain du recensement : innovations en matière de collecte de données par le [1](#), [2](#), [3](#), [4](#)

C

cancer : téléphones portables et [1](#), [2](#), [3](#)

Captcha & ReCaptcha : von Ahn invente [1](#), [2](#)

cartes perforées : Hollerith et [1](#)

cartes perforées :Hollerith et [1](#)

cartographie [1](#)

Cate, Fred [1](#)

causalité : big data et [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#)

Kahneman au sujet de la [1](#), [2](#)
nature de la [1](#), [2](#), [3](#), [4](#)
par opposition à corrélation [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18](#), [19](#), [20](#), [21](#)
préférence intuitive pour [1](#), [2](#), [3](#), [4](#)

Cavallo, Alberto [1](#)

Centers for Disease Control : limites de leur système de suivi de la pandémie [1](#), [2](#), [3](#), [4](#)

CERN laboratoire [1](#)

Chagall, Marc [1](#)

changements climatiques : big data et [1](#)

châtiment : basé sur l'analyse prédictive [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#)

chiffres arabes [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

chiffres romains [1](#), [2](#), [3](#)

« cimetières de données » [1](#)

ClearForest [1](#)

Code for America [1](#)

collecte de données [1](#), [2](#), [3](#)

dans les élections de 2008 [1](#)

dans les sciences sociales [1](#), [2](#)

innovations par le Bureau américain du recensement [1](#), [2](#)

notification et consentement pour la [1](#), [2](#)

par des automobiles [1](#), [2](#), [3](#), [4](#), [5](#)

par des compteurs d'électricité [1](#), [2](#)

par des téléphones portables [1](#), [2](#), [3](#), [4](#)

par la NSA [1](#), [2](#)

pour des recensements [1](#), [2](#), [3](#), [4](#)

s'opposer à l'utilisation de ses [1](#), [2](#)

sociétés axées sur la [1](#), [2](#), [3](#), [4](#), [5](#)

commerce électronique : big data dans le [1](#), [2](#), [3](#)

Commissions de libération conditionnelle : utilisent l'analyse prédictive [1](#)

comportement humain : mise en données et [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

comptage : dans l'Antiquité [1](#), [2](#)

compteurs d'électricité : collecte de données par des [1](#), [2](#)

conception de base de données : exactitude dans la [1](#), [2](#), [3](#)

conception de jeux vidéo : analyse des corrélations dans la [1](#), [2](#), [3](#)

Conférence internationale de Washington (1884) [1](#)

contrôle de qualité : échantillonnage statistique pour le [1](#)

corrélation : *par opposition* à causalité [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18](#), [19](#)

nature de la [1](#), [2](#), [3](#), [4](#)

Coursera [1](#), [2](#)

Craigslist [1](#)
Crawford, Kate [1](#)
Crosby, Alfred [1](#), [2](#), [3](#)
Cross, Bradford [1](#), [2](#), [3](#), [4](#)
culpabilité par association : profilage et [1](#), [2](#)
« culturomics » [1](#), [2](#), [3](#)

D

DataMarket [1](#)
DataSift [1](#)
Davenport, Thomas [1](#), [2](#)
Decide.com [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#)
décisions fondées sur des données [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)
défaillances mécaniques et structurelles : analyse prédictive dans les [1](#), [2](#), [3](#),
[4](#), [5](#), [6](#), [7](#), [8](#)
Delano, Robert [1](#)
Deloitte Consulting [1](#)
Département de la sécurité intérieure des États-Unis [1](#)
 utilise l'analyse prédictive [1](#)
Derawi Biometrics [1](#)
Derwent Capital [1](#)
désordre. *Voir* imprécision
développement économique : big data dans le [1](#), [2](#)
Domesday Book [1](#), [2](#)
données ergonomiques : Koshimizu analyse des [1](#), [2](#), [3](#), [4](#), [5](#)
données ouvertes. *Voir aussi* big data \; données \; informations
 Banque mondiale et [1](#)
 dans l'Union européenne [1](#)
 en Grande-Bretagne [1](#)
 gouvernement et [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)
 nature publique des [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#)
 Obama au sujet des [1](#)
données personnelles : publication par des sociétés de [1](#), [2](#)
 détention de [1](#), [2](#), [3](#)
 protection de la vie privée et [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)
données récupérées de téléphones portables : Apple et [1](#)
 dans le domaine de la santé [1](#), [2](#), [3](#)
 et géolocalisation [1](#), [2](#), [3](#)
 prédissent une épidémie de grippe [1](#)

protection de la vie privée et [1](#), [2](#)

réutilisation de [1](#), [2](#)

données substitutives : dans l'analyse des corrélations [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)

données sur les ventes : analyse des [1](#), [2](#), [3](#), [4](#)

données. *Voir aussi* big data \; informations \; données ouvertes

« dictature » des [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#)

agrégation de [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18](#)

anonymisation de [1](#), [2](#)

comparées à l'énergie [1](#), [2](#)

courtage de [1](#)

coût de stockage des [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

décisions fondées sur des [1](#), [2](#), [3](#), [4](#), [5](#)

dépréciation de la valeur des [1](#), [2](#)

en tant que vérité [1](#), [2](#)

évaluation des [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

exploitation à mauvais escient des [1](#), [2](#), [3](#), [4](#), [5](#)

extensibilité des [1](#), [2](#)

extraction de [1](#), [2](#), [3](#), [4](#), [5](#)

faillibilité des [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#)

fétichisation des [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)

imprécision dans le traitement des [1](#), [2](#)

nature des [1](#), [2](#), [3](#)

recombinaison des [1](#), [2](#), [3](#), [4](#)

taille des [1](#), [2](#), [3](#), [4](#)

valeur d'option des [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)

valeur économique de la réutilisation des [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18](#), [19](#)

dossiers médicaux : analyse des corrélations dans des [1](#), [2](#), [3](#), [4](#)

Dostert, Leon [1](#)

Duhigg, Charles : *Le pouvoir des habitudes* [1](#), [2](#)

E

Eagle, Nathan [1](#), [2](#), [3](#)

eBay [1](#), [2](#)

échantillonnage statistique : les big data remplacent l' [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#)

dans le contrôle de qualité [1](#)

exactitude nécessaire dans l' [1](#), [2](#), [3](#)

Graunt et [1](#), [2](#), [3](#)

limites inhérentes à l' [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

Neyman au sujet de l' [1](#)

sélection aléatoire nécessaire dans l' [1](#), [2](#), [3](#)

Silver au sujet de l' [1](#)

taille dans l' [1](#)

edX [1](#)

Eisenstein, Elizabeth [1](#)

Elbaz, Gil [1](#)

élections de 2008 : collecte de données lors des [1](#)

emplacement géospatial [1](#)

énergie : données comparées à l' [1](#)

enseignement universitaire : utilisation erronée de données dans l' [1](#)
en ligne [1](#), [2](#), [3](#)

Equifax [1](#), [2](#), [3](#)

Eratosthène [1](#), [2](#)

État policier : Allemagne de l'Est en tant qu' [1](#), [2](#), [3](#)

éthique : en matière de big data [1](#), [2](#)

étiquetage : par opposition à mise en catégories [1](#), [2](#)

Etzioni, Oren [1](#), [2](#), [3](#), [4](#)

analyse les fluctuations des prix de billets des compagnies aériennes [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#),
[11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18](#), [19](#)

Euclide [1](#)

Evans, Philip [1](#)

exactitude. *Voir aussi* imprécision

dans la conception de bases de données [1](#), [2](#), [3](#)

et big data [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#)

et mesure [1](#), [2](#)

nécessaire dans l'échantillonnage, [1](#), [2](#), [3](#), [4](#)

Excite [1](#)

Experian [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)

expert scientifiques en données [1](#)

expertise en la matière : rôle dans les big data [1](#), [2](#), [3](#), [4](#), [5](#)

explicabilité : big data et [1](#)

F

Facebook [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#)

mise en données par [1](#), [2](#)

offre publique initiale (OPI) de [1](#), [2](#), [3](#), [4](#), [5](#)

traitement des données par [1](#)

utilise les « traces numériques », [1](#), [2](#)

valorisation de [1](#), [2](#), [3](#)

Factual [1](#)

Fair Isaac Corporation (FICO) [1](#)

Farecast [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#)

films hollywoodiens : recettes futures prédites [1](#), [2](#)

filtrage collaboratif : chez Amazon [1](#), [2](#), [3](#), [4](#)

chez Netflix [1](#)

finance : les big data dans le domaine de la [1](#), [2](#), [3](#), [4](#)

Fitbit [1](#)

Flickr [1](#), [2](#)

FlightCaster.com [1](#), [2](#), [3](#), [4](#), [5](#)

Flowers, Mike : et utilisation de big data par le gouvernement [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

FlyOnTime.us [1](#), [2](#), [3](#), [4](#), [5](#)

Ford Motor Company [1](#), [2](#), [3](#), [4](#)

Ford, Henry [1](#)

Foursquare [1](#), [2](#)

fraude à la carte de crédit : big data et [1](#), [2](#), [3](#), [4](#)

Kunze au sujet de la [1](#)

Freakonomics (Leavitt) [1](#), [2](#)

G

Galton, Sir Francis [1](#)

Gasser, Urs [1](#)

Gates, Bill [1](#)

Géographie (Ptolémée) [1](#)

Géographie physique de la mer (La) (Maury) [1](#)

géolocalisation : données récupérées de téléphones portables et [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

applications commerciales des données [1](#), [2](#), [3](#), [4](#), [5](#)

le secteur des assurances utilise les données de [1](#)

mise en données d'un lieu [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

UPS utilise les données de [1](#), [2](#)

Gnip [1](#)

Goldblum, Anthony [1](#)

Google [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)

analyse des requêtes par [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

cartes [1](#), [2](#)

classement d'un site web par [1](#)

collecte les données GPS [1](#), [2](#), [3](#)

en tant que société des big data [1](#)

et protection de la vie privée [1](#), [2](#)

et traduction de langues [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)

Gmail [1](#), [2](#)

Google Docs [1](#)

intelligence artificielle chez [1](#)

MapReduce [1](#), [2](#)

PageRank [1](#)

prédit une épidémie de grippe [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#)
projet de livres [1](#), [2](#), [3](#), [4](#), [5](#)
reconnaissance vocale chez [1](#), [2](#)
réutilisation de données par [1](#), [2](#), [3](#), [4](#)
Street View, véhicules du programme [1](#), [2](#), [3](#), [4](#)
Suivi de la grippe [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)
système de correction orthographique [1](#), [2](#), [3](#), [4](#)
traitement des données par [1](#)
utilise les « traces numériques » [1](#), [2](#)
utilise les modèles mathématiques [1](#), [2](#), [3](#)
gouvernement : et données ouvertes [1](#), [2](#), [3](#), [4](#), [5](#)
réglementation et big data [1](#), [2](#)
surveillance par la [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)
Grande-Bretagne : données ouvertes en [1](#)
Graunt, John : et échantillonnage [1](#)
grippe : les données récupérées de téléphones portables prédisent l'épidémie de [1](#), [2](#)
Google prédit l'épidémie de [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#)
vaccins [1](#), [2](#)
guerre d'Irak : analyse prédictive durant la [1](#)
guerre du Vietnam : données exploitées à mauvais escient dans la [1](#), [2](#), [3](#), [4](#), [5](#)
Gutenberg, Johannes [1](#)

H

Hadoop [1](#), [2](#)
Hammerbacher, Jeff [1](#)
Harcourt, Bernard [1](#)
Health Care Cost Institute [1](#)
Hellend, Pat : « Si vous avez trop de données, alors “assez bon” est vraiment assez bon » [1](#)
Hilbert, Martin : tentatives de mesure des informations [1](#), [2](#), [3](#)
Hitwise [1](#), [2](#), [3](#), [4](#)
Hollerith, Herman : et cartes perforées [1](#), [2](#)
Homme numérique (L') (Negroponte) [1](#)
Honda [1](#)
Huberman, Bernardo : et analyse des réseaux sociaux [1](#)

I

IBM [1](#)

création de [1](#)
et traduction de langues [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)
et voitures électriques [1](#), [2](#), [3](#), [4](#)
Projet Candide [1](#), [2](#), [3](#)

ID3 [1](#)

immobilier : réglementation relative aux conversions illégales [1](#), [2](#), [3](#), [4](#)

Import.io [1](#)

imprécision. *Voir aussi* exactitude

dans le traitement des données [1](#), [2](#)
en tant que caractère positif des big data [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)
et taille [1](#), [2](#), [3](#), [4](#)
et vérité [1](#)
nature de l' [1](#), [2](#)

indice des prix à la consommation (CPI) [1](#), [2](#)

inflation : big data et calcul de [1](#), [2](#)

information, technologie de l' [1](#), [2](#)

informations. *Voir aussi* big data données \; données ouvertes

analyse des [1](#), [2](#), [3](#)
augmentation du volume d' [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#)
changements qualitatifs des [1](#)
comme base de l'univers [1](#), [2](#)
histoire des [1](#), [2](#), [3](#), [4](#)
innovations dans le traitement des [1](#), [2](#), [3](#)
lois pour l'utilisation des [1](#)
quantité mondiale d' [1](#)
tentatives d'Hilbert pour mesurer les [1](#), [2](#)

Infoseek [1](#)

InfoSpace [1](#)

Inrix : analyse des conditions de circulation par [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)

intelligence artificielle : big data et [1](#)

chez Google [1](#)

intermédiaires en données : sociétés fonctionnant comme des [1](#), [2](#), [3](#), [4](#), [5](#), [6](#),
[7](#), [8](#), [9](#), [10](#), [11](#), [12](#)

Internet : protection de la vie privée et [1](#)

Internet Movie Database [1](#)

intuition : *par opposition à* analyse de données [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)

iPhone [1](#)

ITA Software [1](#), [2](#), [3](#), [4](#), [5](#)

iTrem [1](#)

J

James, Bill [1](#)
Jana [1](#)
Jawbone [1](#)
Jetpac [1](#)
Jobs, Steve [1](#), [2](#)
 et séquençage de l'ADN [1](#), [2](#), [3](#)
Jonas, Jeff [1](#)
justice : fondée sur le libre arbitre [1](#), [2](#), [3](#), [4](#), [5](#)

K

Kaggle [1](#), [2](#), [3](#)
Kahneman, Daniel : au sujet de la causalité [1](#), [2](#)
Kelvin, William Thomson, Lord [1](#)
Kennedy, Len [1](#)
Kennedy, Ted [1](#), [2](#)
Khandelwal, Shashank [1](#), [2](#)
Kindle, liseuse de livres numériques [1](#), [2](#), [3](#), [4](#)
Kinnard, Douglas : *The War Managers* [1](#)
Koshimizu, Shigeomi : analyse des données ergonomiques [1](#), [2](#), [3](#), [4](#), [5](#)
Kunze, John : au sujet de la fraude à la carte de crédit [1](#)

L

Laney, Doug [1](#), [2](#), [3](#)
Large Synoptic Survey Telescope [1](#)
Le Stratège (Lewis) [1](#)
Le Stratège [film] [1](#), [2](#), [3](#), [4](#), [5](#)
Leavitt, Stephen : *Freakonomics* [1](#), [2](#), [3](#)
Levis, Jack [1](#), [2](#), [3](#)
Lewis, Michael : *Le Stratège* [1](#)
lexicologie, par informatique [1](#)
libre arbitre : justice fondée sur le [1](#), [2](#), [3](#), [4](#), [5](#)
 par opposition à analyse prédictive [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)
Linden, Greg [1](#), [2](#), [3](#), [4](#), [5](#)
LinkedIn [1](#), [2](#), [3](#), [4](#), [5](#)
livres numériques. *Voir aussi* livres
 Amazon et [1](#), [2](#)

et mise en données [1](#), [2](#)

et réutilisation de données [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

livres. *Voir aussi* livres numériques

comptes-rendus chez Amazon [1](#), [2](#), [3](#), [4](#)

numérisation & mise en données de [1](#), [2](#), [3](#), [4](#)

lois : contre l'utilisation abusive des big data [1](#), [2](#), [3](#), [4](#)

régissant l'utilisation des informations [1](#)

relatives à la protection de la vie privée [1](#), [2](#)

Luther, Martin [1](#)

lutte sumo : corruption dans la [1](#), [2](#), [3](#)

Lytro, appareil photographique [1](#), [2](#)

M

Marcken, Carl de [1](#), [2](#)

Marcus, James : *Amazonia* [1](#)

MarketPsych [1](#), [2](#)

MasterCard [1](#), [2](#), [3](#), [4](#)

Maury, Matthew Fontaine : *Géographie physique de la mer* [1](#), [2](#)

révolutionne la navigation maritime [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#)

Mayer, Marissa [1](#)

McGregor, Carolyn : et les naissances de prématurés [1](#), [2](#), [3](#)

McKinsey Global Institute [1](#), [2](#)

McNamara, Robert : et analyse de données [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#)

au poste de Secrétaire à la Défense [1](#)

Avec le recul [1](#)

médias électroniques : Prismatic analyse les [1](#), [2](#), [3](#), [4](#)

médias sociaux : mise en données par les [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)

médiateurs [1](#), [2](#)

Medicis, famille [1](#)

MedStar Washington Hospital Center (Washington, D.C.) [1](#), [2](#), [3](#), [4](#)

Mercator, Gerardus [1](#), [2](#)

Merrill, Douglas [1](#), [2](#)

mesure : dans la mise en données [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

exactitude et [1](#), [2](#), [3](#)

MetaCrawler [1](#)

metadonnées : dans la mise en données [1](#), [2](#)

méthode scientifique : *par opposition à* Analyse des corrélations [1](#), [2](#), [3](#), [4](#)

Microsoft [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

et évaluation des données [1](#), [2](#)

et traduction de langues [1](#)
logiciel Amalga [1](#), [2](#), [3](#), [4](#), [5](#)
système de correction orthographique de Word [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

1984 (Orwell) [1](#), [2](#)

Minority Report [film] [1](#), [2](#), [3](#)

Mise en catégories : *par opposition* à étiquetage [1](#), [2](#)

mise en données : de livres [1](#), [2](#), [3](#), [4](#)

d'un emplacement géospatial [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)

de textes [1](#), [2](#)

en tant que projet [1](#), [2](#)

et comportement humain [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

et pointages de crédit [1](#)

livres numériques et [1](#), [2](#)

mesure dans la [1](#)

métadonnées dans la [1](#), [2](#), [3](#)

nature de la [1](#), [2](#), [3](#)

par Facebook [1](#), [2](#)

par les médias sociaux [1](#), [2](#), [3](#), [4](#), [5](#)

par opposition à numérisation [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#)

par Twitter [1](#), [2](#), [3](#), [4](#)

placement en bourse [1](#), [2](#)

revêtement de sol tactile et [1](#)

modèles d'entreprise : big data et [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)

modèles mathématiques : Google utilise des [1](#), [2](#), [3](#)

moteurs de recherche et [1](#), [2](#)

Moore, loi de [1](#), [2](#)

moteurs de recherche : et modèles mathématiques [1](#)

Mydex [1](#)

N

naissances, prématurés : McGregor et [1](#), [2](#), [3](#)

nanotechnologie : et changements qualitatifs [1](#), [2](#)

Nash, Bruce [1](#)

nations : big data et avantage concurrentiel parmi les [1](#), [2](#)

navigation, maritime : analyse des corrélations dans le domaine de la [1](#), [2](#)

Maury révolutionne la [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#)

Negroponte, Nicholas : *L'Homme numérique* [1](#)

Netbot [1](#)

Netflix [1](#)

filtrage collaboratif chez [1](#)

publie des données personnelles [1](#), [2](#), [3](#)

réutilisation de données par [1](#)

New York Times [1](#), [2](#)

New York, Ville de : explosion des plaques en fonte des trous d'homme dans la [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

réutilisation de données du gouvernement dans la [1](#), [2](#), [3](#), [4](#)

Next Jump [1](#)

Neyman, Jerzy : au sujet de l'échantillonnage statistique [1](#)

Ng, Andrew [1](#)

Nippo-américains : internement des (1942) [1](#)

Norvig, Peter [1](#), [2](#)

« L'efficacité déraisonnable des données » [1](#)

Nuance : n'a pas compris l'importance de la réutilisation de données [1](#), [2](#)

numérisation : *par opposition* à mise en données [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)

révolution de la [1](#), [2](#), [3](#)

O

O'Reilly, Tim [1](#)

Oakland Athletics [1](#), [2](#), [3](#)

Obama, Barack : au sujet des données ouvertes [1](#), [2](#)

Och, Franz Josef [1](#)

Ohm, Paul : au sujet de la protection de la vie privée [1](#)

Omidyar, Pierre [1](#)

Open Data Institute [1](#)

Open Knowledge Foundation [1](#)

Organisation internationale de normalisation (ISO) [1](#), [2](#)

Orwell, George : *1984* [1](#), [2](#)

P

Pacioli, Luca : et tenue des livres en partie double [1](#), [2](#), [3](#)

Page, Larry [1](#)

Palfrey, John [1](#)

Parise, Brian [1](#)

Pasteur, Louis : et le vaccin contre la rage [1](#), [2](#), [3](#)

PayPal Mafia [1](#)

Pays-Bas : registres d'état civil très complets aux [1](#), [2](#)

Pearl, Judea [1](#)

Pentland, Sandy [1](#), [2](#)

perceptions humaines : les big data changent les [1](#)
physique quantique [1](#)
Picasso, Pablo [1](#), [2](#)
Pinterest [1](#)
placement en bourse : mise en données dans le [1](#), [2](#)
plaques en fonte de trous d'homme, explosion [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)
pointages de crédit : analyse des corrélations et [1](#)
mise en données et [1](#)
police : utilise l'analyse prédictive [1](#), [2](#), [3](#), [4](#), [5](#)
police prédictive [1](#)
et prévention du crime [1](#), [2](#), [3](#)
Pouvoir des habitudes (Le) (Duhigg) [1](#)
précision. *Voir* exactitude
presse à imprimer : effets socioéconomiques de la [1](#), [2](#), [3](#), [4](#)
prêts personnels : les big data dans les [1](#), [2](#)
prévention du crime : police prédictive et [1](#), [2](#)
prévisions de prix : pour les produits de consommation [1](#), [2](#), [3](#)
PriceStats [1](#)
Prismatic : analyse les médias électroniques [1](#), [2](#), [3](#)
prix des produits de consommation : prévisions relatives aux [1](#), [2](#), [3](#)
profilage : et culpabilité par association [1](#), [2](#)
analyse prédictive dans le [1](#)
progrès : en tant que concept [1](#), [2](#)
Projet Gutenberg [1](#)
protection de la vie privée : et anonymisation [1](#), [2](#), [3](#)
et big data [1](#), [2](#), [3](#), [4](#), [5](#)
et données personnelles [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)
et données récupérées de téléphones portables [1](#), [2](#)
et Internet [1](#)
et notification et consentement [1](#), [2](#), [3](#), [4](#)
et opposition à l'utilisation des données personnelles [1](#), [2](#)
Google et [1](#), [2](#)
lois sur la [1](#), [2](#)
Ohm au sujet de la [1](#)
Ptolémée : *Géographie* [1](#)

Q

Quantcast [1](#), [2](#)
quantification. *Voir* mesure

« quantified self », mouvement [1](#)

R

raffineries de pétrole : big data dans les [1](#)

reality mining (extraction de la réalité) [1](#), [2](#)

recensements : collecte de données pour des [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

reconnaissance vocale : chez Google [1](#)

réglementation antitrust : big data et [1](#), [2](#), [3](#), [4](#), [5](#)

requêtes : analyse et réutilisation des [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)

responsabilité, individuelle : big data et [1](#), [2](#), [3](#), [4](#), [5](#)

Etzioni analyse les fluctuations des prix de billets d'avion [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#)

prédictions de retards de vols [1](#), [2](#), [3](#), [4](#)

Reuters [1](#)

revêtement de sol, tactile : et mise en données [1](#)

Rigobon, Roberto [1](#)

Roadnet Technologies [1](#), [2](#)

Rolls-Royce [1](#), [2](#)

Rudin, Cynthia [1](#), [2](#), [3](#)

Rudin, Ken [1](#)

S

sabermétrie [1](#)

Saddam Hussein : procès de [1](#)

Salathé, Marcel [1](#), [2](#)

Salesforce.com [1](#)

santé : big data dans le domaine de la [1](#), [2](#), [3](#), [4](#), [5](#)

analyse prédictive dans le domaine de la [1](#), [2](#), [3](#), [4](#)

données récupérées de téléphones portables dans le domaine de la [1](#), [2](#), [3](#), [4](#), [5](#)

santé publique : limites du système de suivi de la pandémie [1](#), [2](#), [3](#), [4](#)

scandale des « subprimes » (2009) : analyse des corrélations et [1](#)

sciences sociales : collecte de données dans les [1](#), [2](#)

Scott, James : *Seeing Like a State* [1](#)

Seeing Like a State (Scott) [1](#)

sélection aléatoire : nécessaire dans l'échantillonnage statistique [1](#), [2](#), [3](#)

Sense Networks [1](#), [2](#)

sentiments, analyse des [1](#), [2](#), [3](#), [4](#)

séquençage de l'ADN : big data dans le [1](#)

coût du [1](#), [2](#)

Steve Jobs et [1](#), [2](#), [3](#)

« Si vous avez trop de données, alors “assez bon” est vraiment assez bon »
(Hellend) [1](#)

Silver, Nate : au sujet de l'échantillonnage statistique [1](#)

Skyhook [1](#)

Sloan Digital Sky Survey [1](#)

Smith, Adam [1](#)

« société de l'information » [1](#), [2](#), [3](#), [4](#), [5](#)

« sociétés des big data » [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#)

Society for American Baseball Research [1](#)

sports : analyse prédictive dans le domaine des [1](#), [2](#), [3](#), [4](#), [5](#)

Stasi [1](#), [2](#), [3](#)

statisticiens : forte demande de [1](#), [2](#), [3](#), [4](#)

statistiques : utilisation militaire des [1](#)

Sunlight Foundation [1](#)

Super Crunchers (Ayres) [1](#)

surveillance : par le gouvernement [1](#), [2](#), [3](#), [4](#)

SWIFT : réutilisation de données par [1](#), [2](#)

Système de positionnement global (GPS), satellites du [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

système métrique [1](#)

systèmes de correction orthographique : et réutilisation de données [1](#), [2](#), [3](#), [4](#)

systèmes numériques : histoire des [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)

T

taille : des données [1](#), [2](#)

dans l'échantillonnage statistique [1](#)

fonctions qualitatives de la [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)

imprécision et [1](#), [2](#), [3](#), [4](#)

Taleb, Nassim Nicholas [1](#)

Target : analyse prédictive par [1](#), [2](#), [3](#)

Telefonica Digital Insights [1](#)

téléphones portables : et cancer [1](#), [2](#), [3](#), [4](#)

tenue des livres, en partie double : histoire de [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

Pacioli et [1](#), [2](#)

Teradata [1](#), [2](#), [3](#)

terrorisme : analyse prédictive et [1](#), [2](#), [3](#), [4](#)

texte : analyse des corrélations dans du [1](#), [2](#), [3](#)

mise en données de [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#)

The-Numbers.com : prédit la rentabilité des films hollywoodiens [1](#), [2](#)

théorie des réseaux [1](#)

big data aux [1](#), [2](#)

Thomson Reuters [1](#)

« traces numériques » [1](#), [2](#), [3](#)

traduction automatique. *Voir* traduction de langues

traduction de langues [1](#)

Google et [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

IBM et [1](#), [2](#), [3](#), [4](#)

Microsoft et [1](#)

traitement du langage naturel [1](#)

transactions : analyse de [1](#)

transparence : des algorithmes [1](#)

Transversale universelle de Mercator (UTM), système de projection de la projection [1](#)

Twitter [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)

analyse des messages par [1](#)

en tant que société des big data [1](#), [2](#)

mise en données par [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)

traitement de données par [1](#)

U

U.S. Bureau of Labor Statistics [1](#), [2](#)

U.S. National Security Agency (NSA) : collecte de données par l' [1](#)

Udacity [1](#)

Union européenne : données ouvertes en [1](#)

univers : informations comme base de l' [1](#), [2](#), [3](#)

« Unreasonable Effectiveness of Data, The » (Norvig) [1](#)

UPS : analyse prédictive par [1](#)

utilise les données de géolocalisation [1](#), [2](#)

UPS Logistics Technologies [1](#), [2](#)

V

vaccin contre la rage : Pasteur et le [1](#), [2](#), [3](#)

valeur, économique : big data et création de [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18](#), [19](#), [20](#), [21](#)

de la réutilisation des données [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#), [14](#), [15](#), [16](#), [17](#), [18](#), [19](#), [20](#)

Varian, Hal [1](#), [2](#)

vérité : données en tant que [1](#), [2](#), [3](#)

imprécision et [1](#), [2](#)

23andMe [1](#)

Visa [1](#)

voitures : systèmes d'antivol pour les [1](#)

collecte de données par [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)

voitures électriques : big data et [1](#), [2](#), [3](#)

IBM et [1](#), [2](#)

von Ahn, Luis : invente Captcha & ReCaptcha [1](#), [2](#)

W

Walmart [1](#)

analyse les données sur les ventes [1](#), [2](#), [3](#), [4](#)

innovations dans la commercialisation par [1](#), [2](#), [3](#), [4](#)

War Managers, The (Kinnard) [1](#)

Warden, Pete [1](#)

Watts, Duncan [1](#)

Weinberger, David [1](#)

Wikipedia [1](#)

Windows Azure Marketplace [1](#)

X

Xoom [1](#), [2](#)

Y

Yahoo [1](#), [2](#), [3](#)

YouTube : traitement des données par [1](#)

Z

Zeo [1](#)

ZestFinance [1](#), [2](#), [3](#), [4](#)

Zillow [1](#), [2](#)

Zuckerberg, Mark [1](#), [2](#)

Zynga [1](#), [2](#), [3](#), [4](#), [5](#)